ESTIMATION, MODEL SELECTION, AND RESILIENCE OF POWER-LAW DISTRIBUTIONS

by

Yafei Wei

BS, Jilin University, 2008

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts and Sciences in partial

fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Yafei Wei

It was defended on

Dec 9, 2015

and approved by

Satish Iyengar, PhD, Dietrich School of Arts and Sciences

Henry Block, PhD, Dietrich School of Arts and Sciences

Kehui Chen, PhD, Dietrich School of Arts and Sciences

Abdus Wahed, PhD, Department of Biostatistics

Dissertation Director: Satish Iyengar, PhD, Dietrich School of Arts and Sciences

ACKNOWLEDGEMENT

I would like to express my appreciation and thanks to all people who have helped and accompanied with me during my PhD stage.

Firstly I thank my advisor Dr. Iyengar. He patiently exposed me to various areas of my interest. He inspired research ideas and respected my ideas too. He provided generous financial supports. He devoted direct supervision and work on my thesis. I learned a lot from Dr. Iyengar, for both academic knowledge and life wisdom.

I deeply thank my family. My parents gave me constant care, love, warmth and support. No words can express my gratitude to them. My husband Huichao accompanied with me along the PhD journey, when we shared experience and supported each other. My baby James is a wonderful baby. He is a super gift to me. He motivates me to work hard to be a good mother.

I thank committee members Dr. Block, Dr. Chen, Dr. Wahed, and Dr. Guclu. They provided valuable comments and suggestions on my thesis.

ESTIMATION, MODEL SELECTION, AND RESILIENCE OF POWER-LAW DISTRIBUTIONS

Yafei Wei, PhD

University of Pittsburgh, 2016

This thesis includes a series studies on power-law distribution, which is a widely used distribution in vast areas such as biology, economy, social science and information science. There are three parts in the thesis.

The first part is parameter estimation of power-law distributions. We categorize variants of power-law distributions into six types. We proposed improvements on the estimation for some types, either decreasing bias or standard deviation of the estimates. We also proposed methods for some types if there is no corresponding estimation method yet.

The second part is model selection between non-truncated and truncated power-law distributions. We evaluated both criterion based methods and test based methods on the model selection, by calculating sensitivity and specificity of each method from simulation studies. We also proved some properties of the calculation to extend the result of the simulation study with a particular parameter setting to more general parameter settings.

The third part is exploring resilience of the power-law degree distribution of scale-free networks. We explored how the degree distribution changes if the network receives attacks to lose vertices and corresponding edges under random removal, normal curve removal and high degree removal strategies. We derived the form of expected degree distribution, which is not power law any more even one vertex is removed. We also conducted a simulation study by using goodness of fit test to see the validity of power law, which shows that power law is very resilient for random removal but fragile for high degree removal. We also conducted a simulation study to observe the change of parameters when the goodness of fit test shows that power law is a good fit.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

A power-law distribution is one with probability density function of the form $p(x) = Cx^{-\alpha}$ for $x \geq \gamma$. More generally, this form may be appropriate over a bounded or truncated range $\gamma \geq x \leq \nu$. It arises from the analysis of measurements that do not peak around a typical (mean or median) value. In such situations, measurements often vary over an enormous range, and the distribution has a long tail. An example is the population of towns and cities, which ranges from only 52 to over 8 million according to the 2000 US census (Newman, 2005).

Power-law distributions are found in many research areas such as computer science, information theory, biology, ecology, astronomy, and economics. Examples of data sets that have been shown to be well fit by a power-law distribution within a certain range include the sizes of earthquakes, sizes of computer files, sizes of craters on the moon, frequency of word usage, number of citations, number of hits on web pages, sales of music recordings, and number of species in biological taxa. (Clauset et al., 2009)

There are many studies of power-law distributions (Newman, 2005). People have found interesting mathematical properties of power-law distributions. For example, it is the only distribution that is scale-free: $p(bx) = g(b)p(x)$. It is a heavy-tailed distribution: the majority of top values of x lie in a small proportion of the distribution at the top. For example, the '$80/20$ rule' with 80 $\%$ of the total wealth in a population being in the hands of the richest 20% of the population is a standard application of a power-law or Pareto distribution. There are also studies of statistical inference and tests concerning the power-law distribution, parts of which will be discussed below. Researchers have also tried to find generation mechanisms for the power-law distribution: these mechanisms include random walks, the Yule process, critical phenomena, self-organized criticality, and combinations of exponentials.

In this thesis, we deal with three statistical aspects of power-law distributions. The first part

is on parameter estimation; the second part is on model selection between truncated and non-truncated power-law distributions; and the third part is on the resilience of power-law degree distribution of networks. For each part, we introduce the problems, summarize previous work, and propose our methods and solutions.

2.0 PARAMETER ESTIMATION

2.1 INTRODUCTION

2.1.1 Density type classification

The expression $p(x) = Cx^{-\alpha}$ for $x \geq \gamma$ is the conventional form of the power-law density. However, there are variants based on whether the power law applies in only a subinterval with a truncation above or below or both, and whether there are other distribution types outside the power-law region. Power-law distribution has both continuous form and discrete form too; in this thesis we only consider estimation of the continuous case.

Different variants of the power-law distribution require different parameter estimation schemes and methods. Below, we classify the variants of power-law distribution into six types, and then discuss estimation for each type. For types which already have well understood methods, we list the commonly used methods for them; for others which have unsatisfactory or no method at all, we propose either an improved estimation method or fill in the gap.

The basic form of the power-law distribution is $p(x) = Cx^{-\alpha}, x \geq \gamma$, and is labeled as Type 1: see Figure 1. For this form, the entire range is governed by a power-law distribution. This is the most commonly used form to fit data. For example, it is used to fit people's annual income (Pareto, 1964), number of papers scientists write (Coile, 1977), sizes of earthquakes (Gutenberg and Richter, 1944), and sizes of moon craters (Neukum and Ivanov, 1994). However, sometimes other types of power-law distribution might be more appropriate for such applications (Burroughs and Tebbens, 2001).

In some cases, there is an upper threshold for large values, due to some natural mechanism or observation limit. For example, the size of a forest fire cannot be infinitely large, but is bounded

(by, say, the size of the largest known forest) (Burroughs and Tebbens, 2001). This form is called a power-law distribution with truncation, and we label it as Type 2: see Figure 1 and the corresponding density function in (2.2). It should be noted that sometimes the term "truncated power-law distribution" is used as the power-law distribution with exponential decay: $f(x) = Cx^{-\alpha}e^{-\beta x}$, which is different from the truncation defined here.

In some cases, the density has other distribution types outside the power-law region. We call that an "impure" power-law distribution. Correspondingly, densities with only power-law forms, such as the Type 1 and Type 2 densities, are called "pure" power-law distributions. Those "impure" distributions, where other distribution types appear at the head and the tail has a power-law we call "head-impure" power-law distributions. Such examples are common, since the power law is usually found in the tail. This can also be partly explained by certain generation mechanisms, in which the power-law is formed asymptotically, that is for large values. Depending on whether the power-law tail is truncated or not, there are "head-impure non-truncated" power-law distributions and "head-impure truncated" power-law distributions, which we call Type 3 and Type 4 respectively, shown in Figure 1 and, with density functions given in (2.3) and (2.4).

We call a distribution Type 5 when the power-law distribution fits small values and the tail follows a different distribution. See Figure 1, which is a "tail-impure" power-law distribution, with density function (2.5). However we have omitted estimation for Type 5 in this thesis because there are few applications in literature for it.

Finally, if the power-law distribution is found in an interval, with both the head and tail following other distributions, it is called a "head-tail-impure" power-law distribution and labeled as Type 6: see Figure 1, and its density function in (2.6). Such examples can be found in studies of insurance policies (Beirlant et al., 2015). It should be noted that truncation is different from "tail-impure". Truncation is for cases where there is no other value greater than the upper boundary, while tail-impure means there are still larger values following other distributions.

We summarize these six types in Table 1.

	pure	head-impure	tail-impure	head-tail impure
non-truncated	Type 1	Type 3	Type 5^*	Type 6^*
truncated	Type 2	Type 4		

Table 1: Density Types. Type 1 is "pure non-truncated", Type 2 is "pure truncated", Type 3 is "head-impure non-truncated", Type 4 is "head-impure truncated", Type 5 is "tail-impure", and Type 6 is "head-tail impure".

*Truncation is only for pure tail, therefore Types 5 and 6 are not referred as truncated or not.

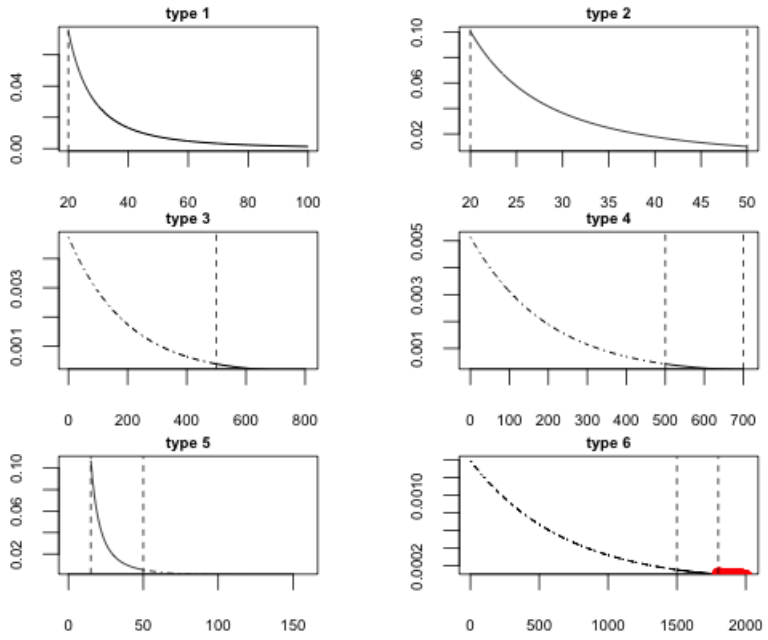

Figure 1: Power-law density variants. The solid curves are power-law parts. The dashed curves (for Type 6 both the dashed curve and the red curve) are not power-law distribution. Vertical dashed lines indicate the borderlines of the power-law part.

The densities for the six types have the following forms:

$$\text{Type 1:} \quad f(x) = Cx^{-\alpha}, \qquad x \geq \gamma \tag{2.1}$$

$$\text{Type 2:} \quad f(x) = Cx^{-\alpha}, \qquad \gamma \leq x \leq \nu \tag{2.2}$$

$$\text{Type 3:} \quad \begin{cases} g(x), & x \leq \gamma \\ f(x) = Cx^{-\alpha}, & x \geq \gamma \end{cases} \tag{2.3}$$

$$\text{Type 4:} \quad \begin{cases} g(x), & x \leq \gamma \\ f(x) = Cx^{-\alpha}, & \gamma \leq x \leq \nu \end{cases} \tag{2.4}$$

$$\text{Type 5:} \quad \begin{cases} f(x) = Cx^{-\alpha}, & \gamma \leq x \leq \nu \\ g(x), & x \geq \nu \end{cases} \tag{2.5}$$

$$\text{Type 6:} \quad \begin{cases} g_1(x), & x \leq \gamma \\ f(x) = Cx^{-\alpha}, & \gamma \leq x \leq \nu \\ g_2(x), & x \geq \nu \end{cases} \tag{2.6}$$

2.1.2 Parameter estimation

Parameter estimation methods are needed for each type of power-law distribution. We only consider the estimation of the power-law part in this thesis, that is, we do not estimate parameters of other distributions outside the power-law region in "impure" cases. In general, parameters that are estimated are boundary parameters γ and ν, and the exponent α (see Eq 2.1 - Eq 2.6).

Typically the distribution parameters will be used in calculating some quantities arising from real problems. For example, if a disease is spreading over a network whose degree distribution follows a power-law distribution, then the speed of its spread is determined by the second moment of the degree distribution (Newman, 2002). The estimation of boundary parameters will affect the estimation of the exponent, which is critical in such calculations. Also, in some cases, the precision of boundary parameters is important in itself. Therefore, we aim to estimate both parameters precisely.

In this thesis, we summarize existing methods for some density types; and propose new methods for other density types if the existing methods are either not satisfactory, or have not yet been developed. Specifically,

1. For Types 1 and 2, we summarize existing methods.
2. For Types 3, 4, and 6, we not only summarize existing methods, but also propose new methods for improvement.

2.2 PREVIOUS WORK

In this section we describe the details of existing estimation methods for each density type.

2.2.1 Type 1

For Type 1, there are two parameters: the lower bound γ and exponent α. Here we introduce the least squares estimate (LSE), a method of moments estimate (MOM), the maximum likelihood estimate (MLE), and quantile estimate (QE). There are other estimates for Type 1 too, but because they are not much used, we omit them: for example, see (Quandt, 1966), which proposes an estimate derived from a type of goodness of fit test.

LSE: The least squares estimate uses a simple power-law property: the log-log plot of the survival function of the power-law distribution should be a straight line, as seen in Figure 2. Specifically, we have

$$p(x) = \frac{\alpha - 1}{1 - \gamma^\alpha} x^{-\alpha} \quad \text{and} \quad S(x) = P(X \geq x) = \frac{1}{1 - \gamma^\alpha} x^{1-\alpha} \quad (x \geq \gamma),$$

so that

$$\ln(S(x)) = -\ln(1 - \gamma^\alpha) + (1 - \alpha)\ln(x);$$

least squares can be used to obtain estimates of γ and α in this linear relationship.

MLE: We start with the likelihood and solve the likelihood equations:

$$\mathcal{L}(\alpha, \gamma | X) = \prod_{i=1}^{n} f(x_i) = (\frac{\alpha - 1}{\gamma^{1-\alpha}})^n \prod_{i=1}^{n} x_i^{-\alpha};$$

$$\frac{\partial \ln \mathcal{L}(\alpha, \gamma | X)}{\partial \alpha} = n[\ln(\alpha - 1) - (1 - \alpha)\ln\gamma] - \alpha \sum_{i=1}^{n} \ln x_i;$$

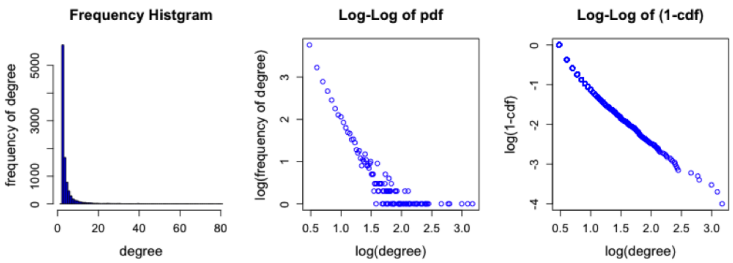

Figure 2: Example of a power-law distribution. The figure shows plots from a network with power-law degree distribution, which includes the histogram of degrees, the empirical pdf of degrees on a log-log scale and the empirical survival density of degrees on a log-log scale.

so that

$$\hat{\alpha}_{mle} = 1 + \frac{n}{\sum_{i=1}^{n} \ln \frac{x_i}{\gamma}} \quad \text{and} \quad \hat{\gamma}_{mle} = X_{(1)}.$$

MOM: We use first moment of an observation and from the first order statistic of a sample from this distribution. If $\alpha > 2$, the mean of power-law distribution exists. Thus, we have

$$E(X) = \frac{\alpha - 1}{\alpha - 2}\gamma \quad \text{and} \quad E(X_{(1)}) = \frac{n(\alpha - 1)\gamma}{n(\alpha - 1) - 1},$$

so that the MOM estimators are

$$\hat{\alpha}_{MOM} = 1 + \frac{n\overline{X} - X_{(1)}}{n(\overline{X} - X_{(1)})} \quad \text{and} \quad \hat{\gamma}_{MOM} = X_{(1)}\left[1 - \frac{\overline{X} - X_{(1)}}{n\overline{X} - X_{(1)}}\right].$$

QE: Choose two probability levels P_1 and P_2, and determine two quantiles x_1 and x_2, that is, $P(X \leq x_1) = P_1$, and $P(X \leq x_2) = P_2$, then

$$P_1 = 1 - \left(\frac{x_1}{\gamma}\right)^{1-\alpha} \quad \text{and} \quad P_2 = 1 - \left(\frac{x_2}{\gamma}\right)^{1-\alpha},$$

so that

$$\hat{\alpha}_{QE} = 1 - \frac{\log \frac{1-P_1}{1-P_2}}{\log \frac{x_1}{x_2}};$$

substituting $\hat{\alpha}_{qtl}$ to one of the above yields $\hat{\gamma}_{QE}$.

(Quandt, 1966) has shown that all estimates above are consistent. He has conducted simulations to compare their performance. No great differences were found between the four methods,

although the MLE and QE (with the 1st quantile) performed best. Comparisons of the mean squared error (MSE) using simulations showed that MLE performs better (Goldstein et al., 2004), (Bauke, 2007), (White et al., 2008).

2.2.2 Type 2

The density function has the form

$$f(x) = Cx^{-\alpha}, \qquad \gamma \leq x \leq \nu.$$

For Type 2, there are several methods in the literature: binning (Apellániz and Úbeda, 2005), CDF (Koen, 2006), minimum variance unbiased estimate (UMVUE) (Beg, 1983), MLE (Aban et al., 2006), improved Hill estimator (same as MLE) (Nuyts, 2010), and bias-free estimators based on the MLE (Maschberger and Kroupa, 2009). Here we describe the MLE, which is most often used. Our methods below are based on it.

MLE: The value of the normalizing constant is

$$C = \frac{\alpha - 1}{\gamma^{1-\alpha} - \nu^{1-\alpha}}.$$

Thus, the likelihood function is:

$$\mathcal{L}(\alpha, \gamma | X) = \prod_{i=1}^{n} f(x_i) = (\frac{\alpha - 1}{\gamma^{1-\alpha} - \nu^{1-\alpha}})^n \prod_{i=1}^{n} x_i^{-\alpha} I(\gamma < X_{(1)} < X_{(n)} < \nu).$$

The partial derivative with respect to α is

$$\begin{aligned} \frac{\partial \ln \mathcal{L}(\alpha, \gamma, \nu | X)}{\partial \alpha} &= \frac{n}{\alpha - 1} + n \frac{\gamma^{1-\alpha} \ln \gamma - \nu^{1-\alpha} \ln \nu}{\gamma^{1-\alpha} - \nu^{1-\alpha}} - \sum_{i=1}^{n} \ln x_i \\ &= \frac{n}{\alpha - 1} + n \frac{\ln \frac{\nu}{\gamma}}{(\frac{\nu}{\gamma})^{1-\alpha} - 1} - \sum_{i=1}^{n} \ln \frac{x_i}{\gamma}. \end{aligned} \tag{2.7}$$

The MLE estimates of the boundaries are

$$\hat{\gamma}_{mle} = X_{(1)} \quad \text{and} \quad \hat{\nu}_{mle} = X_{(n)};$$

substituting $\hat{\gamma}_{mle}$ and $\hat{\nu}_{mle}$ into (2.7) yields $\hat{\alpha}_{mle}$ using a one-dimensional search. (Aban et al., 2006) has proved consistency and asymptotic normality of MLE of α.

2.2.3 Type 3

The density function is:

$$\begin{cases} g(x), & x \leq \gamma \\ f(x) = Cx^{-\alpha}, & x \geq \gamma \end{cases}$$

If γ is known, Type 3 is similar to Type 1, but easier because there is only one unknown parameter α to estimate. We can immediately get the MLE, a moment estimate, the LSE, and quantile estimate by reference of the formulas in Type 1. Hill's estimator (1975) is equivalent to the MLE when γ is known. In certain application areas other methods have been proposed; for example, in the finance and insurance industries, robust estimate and shrinkage estimate have been proposed by (Brazauskas and Serfling, 2003), (Singh et al., 2007). Hill's estimator is based on order statistics: suppose that $X_{(1)} \geq X_{(2)} \geq ... \geq X_{(k)} ... \geq X_{(n)}$, and it is known that $X(k) \geq \gamma$, then

$$\hat{\alpha}_{Hill} = \frac{1}{k} \sum_{i=1}^{k} (\ln X_{(1)} - \ln X_{(k)})$$

If γ is unknown, it is necessary to first estimate γ, and then estimate α with that estimate. One way to detect γ manually is to try many values of γ to estimate α, and to choose the value of γ for which the estimate of α becomes stable. Two shortcomings of this process are that it is laborious and not precise (White et al., 2008). (Clauset et al., 2009) proposed the following approach to estimate γ automatically. For each given γ, fit the data above that γ by a power-law density and use MLE to estimate α for that part of data; then calculate the goodness of fit test statistic KS by using Kolmogorov-Smirnov (KS) test for that part of data; choose the value of γ which yields the smallest KS statistic.

In our work below, we find that, in some cases the bias and standard deviation of the Clauset method are not small. We attempt to decrease both the bias and standard deviation.

2.2.4 Type 4

Density function is:

$$\begin{cases} g(x), & x \leq \gamma \\ f(x) = Cx^{-\alpha}, & \gamma \leq x \leq \nu \end{cases}$$

If γ is known, Type 4 is similar to Type 2, but easier because there are only two unknown parameters α and ν. Both the MLE and UMVUE are available here (Beg, 1983).

If γ is unknown, we need to search for the appropriate γ first. Just as for Type 3, (Aban et al., 2006) has proposed the examination of different values of γ to choose the one for which the MLE of α becomes stable. However, there is no work on an automatic search of γ yet. In this thesis, we adapt the automatic search of γ in Type 3 to this framework.

2.2.5 Type 6

The density function is:

$$\begin{cases} g_1(x), & x \leq \gamma \\ f(x) = Cx^{-\alpha}, & \gamma \leq x \leq \nu \\ g_2(x), & x \geq \nu \end{cases}$$

For this type, both endpoints must be estimated. (Deluca and Corral, 2013) and (Peters et al., 2010) have proposed a method for searching γ and ν by modifying Clauset method. Instead of searching one boundary parameter γ in Clauset method, they proposed to search two parameters γ and ν together. For each potential pair of the parameters, calculate the p-value of the KS test for the data bracketed by this pair of ends. They select parameter pairs whose p-values are beyond a threshold, and choose the parameter pair whose interval is longest. They applied their method to real data without assessing its performance using simulation studies. Here we conduct a simulation study to study its performance. We also present methods for comparison.

2.3 OUR PROPOSAL FOR TYPES 3, 4, AND 6

2.3.1 Type 3

2.3.1.1 Examination of Clauset method We examine the performance of Clauset's method (Clauset et al., 2009) for comparatively small data sets using simulations. We explored more parameter combinations of n, γ, and α than they did. In (Clauset et al., 2009), the simulation sample size was 50,000; however, in many applications of the power-law distribution tails, data

are not that rich. For example, the power-law distribution has been applied to the analysis of large US weather losses between 1980 to 2011, which has only 36 data points: (Clark, 2013). In this thesis, we used sample sizes of 100, 500 and 1000 in simulations. In the simulation of Clauset paper, $\alpha = 2.5$. In this thesis we tried other values: 1.5, 2, 2.5, and 3, where $\alpha = 2$ is the boundary determining whether the mean of the power-law distribution exists or not.

We only used $\gamma = 15$ in this thesis because we prove in Lemma 1 below that for the power-law distribution from which data is simulated, result of one value of γ is enough to infer results of all of values. We simulated data sets with the form (2.8) that is used in (Clauset et al., 2009). This distribution has the features that it has continuous derivatives at γ, making estimation more challenging; furthermore, it will guarantee the effectiveness of an estimation method if it performs well in simulations. There is another important property of this distribution, which is proved in Lemma 1: after data is multiplied by a constant, the new data will still follow this distribution, with the same α, and with γ multiplied by the same constant. By this property, in Lemma 1 it proves that if n and α fixed, the bias and standard deviation of estimates of α (denoted as $\hat{\alpha}$) will be the same for different γ values; the bias and standard deviation of estimates of γ (denoted as $\hat{\gamma}$) will be proportional to the value of γ. It follows that only one value of γ is enough in this simulation. To illustrate this, suppose n and α are fixed, we compare estimates for $\gamma = 15$ and $\gamma = 150$. Using Clauset method for estimation, bias and standard deviation of $\hat{\alpha}$ are same for $\gamma = 15$ and $\gamma = 150$; bias and standard deviation of $\hat{\gamma}$ for $\gamma = 150$ is 10 times of those for $\gamma = 150$.

$$p(x) = \begin{cases} C(x/\gamma)^{-\alpha}, \ for\ x \geq \gamma \\ Ce^{-\alpha(x/\gamma - 1)}, \ for\ x < \gamma \end{cases} \tag{2.8}$$

Lemma 1 is valid not only for the Clauset method, but also for methods that were introduced later, such as the jackknife, Hall's bootstrap, Danielsson, KS+rank, and Hall+Clauset. The extensions can be proved the same way.

Lemma 1: Suppose X follows power-law distribution in the form Eq 2.8, with parameters α_X and γ_X; Y follows power-law distribution in the form Eq 2.8 too, with $\alpha_Y = \alpha_X$, and $\gamma_Y = m\gamma_X$, where m is a positive constant. Denote bias as b and variance as S. Then if using Clauset method to do estimation for γ and α, we have: $b(\hat{\alpha}_X) = b(\hat{\alpha}_Y), \quad S(\hat{\alpha}_X) = S(\hat{\alpha}_Y), \quad b(\hat{\gamma}_Y) = mb(\hat{\gamma}_X), \quad S(\hat{\gamma}_Y) = m^2 S(\hat{\gamma}_X)$.

Proof: Using Clauset method, $\hat{\gamma_X} = \underset{\gamma}{\operatorname{argmin}}\ KS(X \geq \gamma) = \underset{\gamma}{\operatorname{argmin}} \max_{x \geq \gamma} |\hat{P}(x) - P(x)|$, where KS is the Kolmogorov-Smirnov test statistic, $\hat{P}(X)$ is the empirical survival function of the part of data which is greater than γ, and $P(X)$ is the theoretical power-law survival function for the same part of data. Similarly, $\hat{\gamma_Y} = \underset{\gamma}{\operatorname{argmin}}\ KS(Y \geq \gamma) = \underset{\gamma}{\operatorname{argmin}} \max_{y \geq \gamma} |\hat{P}(y) - P(y)|$.

For $m > 0$, the density of mX is

$$p(x) = \begin{cases} C(x/m\gamma)^{-\alpha}, \ \text{ for } \ x \geq m\gamma \\ p(x) = Ce^{-\alpha(x/m\gamma - 1)}, \ \text{ for } \ x < m\gamma, \end{cases} \tag{2.9}$$

so that Y and mX have the same distribution. We can suppose there exists a one-to-one mapping from the sample space of X to the sample space of Y: for any given sample $x_1, x_2, ...x_n$ from X, there exists a sample from Y which is $y_1 = mx_1, y_2 = mx_2, ..., y_n = mx_n$, and vice versa for any given sample from Y.

Consider any sample $\{x_1, x_2, ..., x_n\}$, and corresponding mapped sample $\{y_1, y_2, ..., y_n\}$. For any given value $x_\star \in \{x_1, x_2, ..., x_n\}$, the mapped value of $\{y_1, y_2, ..., y_n\}$ is $y_\star = mx_\star$. $\hat{P}(y_\star) = \hat{P}(x_\star)$, since the value of empirical survival function is the percentage of data greater than the lower bound, and these two samples are proportional, with proportional lower bounds for power-law distribution too. $P(y_\star) = \int_{y_\star}^{\infty} p_Y(t)\, dt = \int_{mx_\star}^{\infty} \frac{1}{m} p_X(\frac{t}{m})\, dt = \int_{x_\star}^{\infty} p_X(t)\, dt = P(x_\star)$. Therefore $\hat{\gamma_Y} = \underset{\gamma}{\operatorname{argmin}} \max_{y \geq \gamma} |\hat{P}(y) - P(y)| = \underset{\gamma}{\operatorname{argmin}} \max_{mx \geq \gamma} |\hat{P}(x) - P(x)| = \underset{\gamma}{\operatorname{argmin}} \max_{x \geq \frac{\gamma}{m}} |\hat{P}(x) - P(x)| = m\hat{\gamma_X}$. $\hat{\alpha}_Y = 1 + \frac{n}{\sum_i \log y_i - n \log y_{(1)}} = 1 + \frac{n}{\sum_i \log mx_i - n \log mx_{(1)}} = 1 + \frac{n}{\sum_i \log x_i - n \log x_{(1)}} = \hat{\alpha}_X$. Since these relationships hold for any sample from X, it follows that

$$b(\hat{\alpha}_Y) = E(\hat{\alpha}_Y - \alpha_Y) = E(\hat{\alpha}_X - \alpha_X) = b(\hat{\alpha}_X)$$

$$S(\hat{\alpha}_Y) = E(\hat{\alpha}_Y)^2 - 2E(\hat{\alpha}_Y)\alpha_Y + (\alpha_Y)^2 = E(\hat{\alpha}_X)^2 - 2E(\hat{\alpha}_X)\alpha_X + (\alpha_X)^2 = S(\hat{\alpha}_X)$$

$$b(\hat{\gamma}_Y) = E(\hat{\gamma}_Y - \gamma_Y) = E(m\hat{\gamma}_X - m\gamma_X) = mb(\hat{\gamma}_X)$$

$$S(\hat{\gamma}_Y) = E(\hat{\gamma}_Y)^2 - 2E(\hat{\gamma}_Y)\gamma_Y + (\gamma_Y)^2 = m^2 E(\hat{\gamma}_X)^2 - 2m^2 E(\hat{\gamma}_X)\gamma_X + m^2(\gamma_X)^2 = m^2 S(\hat{\gamma}_X).$$

Clauset's method is based on the assumption that if the candidate γ is bigger than the true γ, the calculated candidate KS will be bigger than KS calculated from true γ because of the worse power-law fitting due to the lack of data; if the candidate γ is smaller than the true γ, then the calculated KS will also be greater than the true KS because of the worse fitting due to adding

impure data from distributions other than the power-law distribution. Therefore, during searching of γ, say searching from the largest to smallest value (though any search direction in principle reaches the same result), Clauset et al. expect KS to decrease first to hit the lowest point when true γ is searched, and then increase again when involving more impure data.

However, this basic assumption is not correct. There are two competing factors when including more impure data (i.e., candidate γ smaller than true γ to include data from other distributions): one is increasing the amount of data which decreases KS, and the other is making the power-law fit worse which increases KS. Therefore, when searching γ from large to small, after hitting true γ and searchomg for smaller γ, encompassing more impure data may cause KS to keep decreasing, because the increase of KS due to the impurity of data cannot overcome the decrease of KS due to the larger data amount. Therefore, it is likely that Clauset method produces an estimate of γ that is biased low.

We examined Clauset estimation with the results shown in Table 2.

The lower bias of the estimate of γ (denoted as $\hat{\gamma}_C$) is apparent, except when $n = 1000$ and $\alpha = 1.5$, however in that case the standard deviation is big to make the marginally higher biased γ untrustworthy. It is concerning that for $n = 1000$ and $\alpha = 2$, the bias of $\hat{\gamma}_C$ is nearly one third of its true values. As we proved in Lemma 1, this bias proportion applies to all values of γ. If $\gamma = 1500$, bias will be around 500. Therefore, correcting the bias of the estimate of γ is necessary.

The bias and standard deviation of $\hat{\alpha}_C$ increase as n becomes small and α becomes large, which is reasonable, because smaller n or larger α both make the amount of data in the power-law tail smaller. The bias of $\hat{\gamma}_C$ changes in the same way, while its standard deviation of moves in the opposite direction. It is possibly because when power-law tail part is small, the search for γ will include values much smaller than the true γ, and searched results scatter within the "impure" region, which is comparatively stable; when the power-law tail part is large, γ will be searched closer to true γ, therefore combating factors mentioned above (increasing data amount and including more impure data) makes searched results jump in and out of the "impure" region, causing a large standard deviation of $\hat{\gamma}_C$. The high standard deviation of γ, for example when $\alpha = 1.5$ is also a concern.

Thus, we aim to decrease bias of $\hat{\gamma}_C$ especially when α is small, and decrease standard deviation especially when α is large. In fact, among the methods below, correcting bias or decreasing

$n = 100$	$\alpha = 1.5$	time	$\alpha = 2$	time	$\alpha = 2.5$	time	$\alpha = 3$	time
Clauset	1.506±0.098 13.757 ± 22.953	1s	1.994 ± 0.233 7.800± 4.253	1s	2.445± 0.437 6.879±3.070	1s	2.789±0.614 6.124± 2.510	1s
Jackknife	1.509±0.083 14.820±20.489		2.002±0.203 7.955±4.392		2.439±0.364 6.833±2.671		2.784±0.523 6.095±2.153	
Danielsson	2.649±9.992 237.764±846.895	4min1s	5.974±23.670 31.506±44.241	4min7s	8.851 ±35.640 18.331±12.787	4min8s	11.882±47.520 15.616±6.709	4min7s
KS+rank	1.487±0.070 7.565 ±8.151	11s	1.927±0.192 6.039±2.791	13s	2.257±0.320 5.198±1.928	13s	2.507± 0.425 4.710±1.743	13s
Clauset+Hall	1.623 ±1.130 54.629±398.051	2min39s	2.253±2.266 11.924 ±13.081	2min4s	2.848±3.414 9.752 ±5.260	1min41s	3.385±4.568 8.753±3.502	1min32s
$n = 500$	$\alpha = 1.5$	time	$\alpha = 2$	time	$\alpha = 2.5$	time	$\alpha = 3$	
Clauset	1.498±0.036 12.404 ±15.109	5s	1.990±0.086 9.904±4.450	5s	2.467±0.161 9.592±2.958	6s	2.882±0.252 8.462±2.2292	5s
Jackknife	1.498±0.035 12.711±14.248		1.990 ± 0.085 9.984±4.337		2.466±0.159 9.548 ±2.811		2.882±0.250 8.482±2.180	
Danielsson	1.492±0.051 12.312±8.435	1hr12min	2.049 ± 0.605 17.182 ±39.966	1hr13min	2.563±0.915 17.469±10.317	1hr12min	3.833±7.323 16.672±7.738	1hr12min
KS+rank	1.495±0.032 8.524±3.313	1min44s	1.968±0.073 8.220± 2.089	1min43s	2.411 ±0.142 7.990±1.780	103s 1min43s	2.787±0.242 7.239 ±1.607	1min43s
Clauset+Hall	1.503±0.032 13.800±8.259	17min19s	1.996±0.082 12.023±3.257	10min15s	2.477 ±0.154 11.668±2.515	6min29s	2.912± 0.246 10.527±2.085	5min13s
$n = 1000$	$\alpha = 1.5$	time	$\alpha = 2$	time	$\alpha = 2.5$	time	$\alpha = 3$	
Clauset	1.501±0.025 16.203±18.791	14s	1.994 ±0.058 10.750±3.970	14s	2.469 ±0.104 10.097±2.772,	15s	2.933± 0.180 9.489± 2.195	14s
Jackknife	1.498±0.028 15.420±16.634		1.986±0.070 11.176±7.424		2.466±0.123 10.146±3.052		2.937±0.211 9.650±2.228	
Danielsson	1.681±1.306 4147.061±23986.95	4hr50min	2.523±2.960 58.744 ±210.826	4hr52min	3.061±3.919 19.556±22.515	4hr53min	4.398±6.480 19.541±13.704	4hr50min
KS+rank	1.497±0.022 9.619±4.526	4min52s	1.985±0.058 9.337 ±2.283	4min50s	2.441 ±0.103 8.780±1.600	4min37s	2.848±0.161 8.132±1.444	4min37s
Clauset+Hall	1.502±0.022 15.203±6.309	38min5s	2.000±0.060 13.359±3.0780	21min34s	2.489±0.100 12.675±1.965	13min22s	2.960±0.181 11.916±2.157	8min55s

Table 2: Various methods for Type 3. True $\gamma = 15$ for all situations. For each parameter combination, 100 data sets are simulated for estimation. For each method, top line is an estimate of α, and bottom line is an estimate of γ and computation time.

standard deviation usually will apply for all α, not specifically for some values of α. While some methods can decrease both bias and standard deviation, some methods can decrease one and increase the other in compensation (a form of bias-variance tradeoff).

2.3.1.2 Possible solutions We list below several methods to correct bias and decrease standard deviation of $\hat{\gamma}_C$. We also examined the estimation on $\hat{\alpha}_C$ too without treating it as our main concern, partly because $\hat{\alpha}_C$ is comparatively similar across methods.

- We tried the jackknife correct for bias, because it is classical bias correction method.
- We adapted bootstrap method proposed by Hall (Hall, 1990) to estimate γ here. It has been proved that such estimate of γ (denoted as $\hat{\gamma}_H$) makes estimate of α (denoted as $\hat{\alpha}_H$) achieve asymptotic minimized MSE. We tried this method by considering that $\hat{\gamma}_H$ which yields a good estimate of $\hat{\alpha}_H$ should be a good estimate for γ itself too.
- We adapted the Danielsson method here (Danielsson et al., 2001). It automatically selects a subsample which is a prerequisite parameter for Hall's method.
- We proposed a new method with the name KS+rank, which uses the spirit of bootstrap but does not require prerequisite subsample size.
- We also proposed a new method with the name Hall+Clauset by combining the Clauset and Hall methods. This method was based on our careful examination of our simulation results.

2.3.1.3 Jackknife The jackknife leaves one data point out and uses Clauset estimation for the remaining data for each time, then the averages of all these estimates to get the final estimate. Table 2 shows that both bias and standard deviation of both $\hat{\gamma}_C$ and $\hat{\alpha}_C$ are improved by the jackknife; however, the improvement seems trivial, and it takes nearly n times longer of computation than Clauset method. It is known that jackknife is useful to correct the bias of estimator if the estimator is smooth; however, γ is not a smooth estimator (it is a boundary point), which is likely the reason why the jackknife is not very effective here.

2.3.1.4 Hall method Hall (Hall, 1990) proposed the use of the bootstrap to estimate MSE; based on it he proposed a parameter estimation method when the purpose of estimation is to minimize the MSE. Unlike searching values of γ in Clauset's method, it searches for which order

statistic can be treated as γ. With such estimated order k it uses the Hill estimator (Hill et al., 1975) to estimate α for the power-law distribution part. This method reaches the asymptotic minimized MSE of estimate of α ($\hat{\alpha}_H$). Hall's method is applicable for more generalized types of distributions: $1 - F(x) \sim Cx^{-\alpha}$ with $C, \alpha > 0$. There are no simulations, for power-law or other distributions in his paper. Also, it is not known that whether $\hat{\gamma}_H$, at least for power-law distributions, will be satisfactory too. Therefore, we explored Hall's estimation for power-law distribution by the simulation.

The procedure consists of the following steps:

- Step 1: Suppose data set is $x_1, x_2, ..., x_n$. Set subsample size n_1.
- Step 2: Search all possible values of k_1 ($k_1 = 2, 3, ..., n_1$). k_1 is a rank such that largest k_1 data in subsample is considered to follow power-law distribution. For each k_1, use Hall's formula to calculate the MSE of the estimate of α of the subsample.
- Step 3: Choose k_1 whose MSE of the estimate of α of the subsample is minimized. Finally, set $k = k_1(\frac{n_1}{n})^{\frac{2}{3}}$ (he gives a theoretical justification for the exponent 2/3); k is the rank such that largest k data of the original data set is considered to follow power-law distribution. Therefore, given $x_{(n)} \leq x_{(n-1)} \leq ... \leq x_1$, $x_{(k)}$ is an estimate for γ.
- However, the formula provided to calculate MSE in Step 2 did not yield reasonable results in our simulations; therefore, we used simulated subsamples to calculate the estimated MSE to replace the theoretical MSE in step 2. So our procedure is:
- Step 1: Suppose data set is $x_1, x_2, ..., x_n$. Set subsample size n_1 and number of subsamples m. Simulate m subsamples.
- Step 2: Search all possible values of rank k_1 ($k_1 = 2, 3, ..., n_1$). For any given value of k_1, for each subsample, treat the largest k_1 data in the subsample as power-law distribution and use maximum likelihood to get $\hat{\alpha}_{k_1,i}, i = 1, 2, ..., m$. Also, use $k = k_1(\frac{n_1}{n})^{\frac{2}{3}}$ to determine the rank in the original data, and get $\hat{\alpha}_k$. Estimate the MSE thus:

$$\hat{MSE}_{k_1} = \sqrt{\sum_{i=1}^{m}(\hat{\alpha}_{k_1,i} - \hat{\alpha}_k)^2/(m-1)}.$$

- Step 3: Choose k_1 which minimizes $\hat{MSE}_{k_1}$. Finally, $k = k_1(\frac{n_1}{n})^{\frac{2}{3}}$. Therefore, given $x_{(n)} \leq x_{(n-1)} \leq ... \leq x_{(1)}$, $x_{(k)}$ is $\hat{\gamma}_H$.

Table 3 shows that different choices of n_1 make a difference in both $\hat{\gamma}_H$ and $\hat{\alpha}_H$. When n_1 is chosen appropriately, Hall's method clearly outperforms Clauset's method, for both bias and standard deviation, and for both α and γ. However if n_1 is not chosen well, performance of Hall's method can be worse than Clauset's. From our simulation study, we found that when n_1/n is chosen around to be the true proportion of power-law tail, Hall estimation performs comparatively best. This rule of thumb appears to be new; we have not seen it in the literature.

When doing a simulation study, it is necessary to decide how many subsamples are needed for calculating the estimated MSE. Table 4 shows that the number of subsamples does not affect the result too much. Therefore we just chose 200 to be the number of subsamples. Using 200 subsamples means that computation time of Hall's method can be controlled at less than 200 times of that of Clauset's method, because length of each subsample is no larger than the original data size.

2.3.1.5 Danielsson method Danielsson (Danielsson et al., 2001) proposed a two-step bootstrap method for determining fraction of the power-law distribution tail part (as in Hall's method, the aim is to determine which order statistic will be $\hat{\gamma}_H$). It is an improvement on Hall's method in that it does not require the important tuning parameter n_1. It automatically searches n_1 by having even smaller subsamples with size n_2.

The steps for Danielsson's method are the following.

- Step 1: Suppose data set is $x_1, x_2, ..., x_n$.
- Step 2: For each n_1 ($n_1 = 2, 3, ..., n$), calculate asymptotic MSE at each k_1 by using simulated subsample as in Hall's method. Here k_1 has the same meaning as in Hall's method - the largest k_1 data of the subsample is considered to follow power-law distribution. Find $k_{1,0}$ for n_1 which minimizes this bootstrap AMSE.
- Step 2: Repeat Step 1 for an even smaller subsample size $n_2 = (n_1)^2/n$, and get $k_{2,0}$ for n_2 the same way.
- Step 3: Choose n_1 which minimizes $\frac{(Q(n_1,k_{1,0}))^2}{Q(n_2,k_{2,0})}$, where Q is asymptotic MSE. Note that a grid search for n_1 is suggested rather than a search for all values of n_1, which will take a lot of time.

$n = 100$	$\alpha = 1.5$	time	$\alpha = 2$	time	$\alpha = 2.5$	time	$\alpha = 3$	time
Clauset	1.51±0.10 13.76 ± 22.95	1s	1.99 ± 0.23 7.80± 4.25	1s	2.44± 0.44 6.88±3.07	1s	2.79±0.61 6.12± 2.51	1s
$0.1n$	1.53±0.11 52.09 ±39.79	11s	2.10±0.28 20.32 ±7.35	12s	2.65 ± 0.46 14.41 ±3.65	12s	3.27±1.09 11.91 ± 2.63	11s
$0.2n$	1.52±0.08 31.73± 22.52	25s	2.06± 0.24 15.17±5.67	25s	2.57±0.47 11.44 ±3.09	25s	3.14±1.47 9.80±2.88	24s
$0.5n$	1.51±0.07 13.75±6.88	1min9s	2.23±2.26 11.06±12.37	1min7s	2.41±0.39 8.05±2.87	1min8s	3.38±4.87 7.39±3.22	1min10s
0.46, 0.24 0.13, 0.07	1.51±0.07 15.27± 7.56	1min3s	2.05±0.26 13.81±6.40	30s	2.62± 0.52 13.41±3.78	15s	3.42±1.16 13.37± 3.51	7s
$n = 500$	$\alpha = 1.5$	time	$\alpha = 2$	time	$\alpha = 2.5$	time	$\alpha = 3$	time
Clauset	1.50±0.04 12.40±15.11	5s	1.99±0.09 9.90±4.45	5s	2.47±0.16 9.59±2.96	6s	2.88±0.25 8.46±2.23	5s
$0.1n$	1.50± 0.04 41.24±22.99	1min8s	2.01 ± 0.09 18.49 ± 3.53	1min10s	2.50±0.16 14.40 ± 2.36	1min7s	2.94±0.23 12.12 ± 1.93	1min7s
$0.2n$	1.50±0.03 25.98±8.31	2min30s	2.00 ±0.08 14.64±2.98	2min30s	2.4953505 ± 0.16 12.48±2.21	2min25s	2.93± 0.24 10.83±1.72	2min25s
$0.5n$	1.50±0.03 14.28 ±3.97	7min25s	1.98 ± 0.07 10.47±2.32	7min19s	2.40±0.15 8.79±1.91	7min19s	2.77±0.23 8.01±1.61	7min22s
0.46, 0.24 0.13, 0.07	1.50± 0.03 15.01±3.50	6min43s	2.00 ± 0.08 14.06±3.15	2min58s	2.49± 0.16 13.62±2.43	1min29s	2.96 ±0.24 12.92±1.73	46s
$n = 1000$	$\alpha = 1.5$	time	$\alpha = 2$	time	$\alpha = 2.5$	time	$\alpha = 3$	
Clauset	1.50±0.02 16.20±18.79	14s	1.99 ±0.06 10.75±3.97	14s	2.47 ±0.10 10.10±2.77	15s	2.93± 0.18 9.49± 2.20	14s
$0.1n$	1.50±0.03 41.10±11.35	2min28s	2.01±0.06 19.45±3.66	2min27s	2.49 ±0.10 14.70±2.02	2min52s	2.97±0.18 12.98± 2.03	2min33s
$0.2n$	1.50±0.03 26.54 ± 7.98	5min35s	2.00 ±0.06 15.71±2.79	5min35s	2.49±0.10 13.08 ±1.67	6min18s	2.95±0.17 11.53±1.77	5min32s
$0.5n$	1.50±0.02 15.03±3.55	19min24s	1.99±0.06 11.31±2.00	19min20s	2.4±0.11 9.76 ± 1.53	21min6s	2.84±0.17 8.94± 1.39	19min21s
0.46, 0.24 0.13, 0.07	1.50±0.02 15.99 ±3.63	17min9s	2.00±0.06 14.75±2.45	7min36s	2.49±0.10 14.06±1.87	3min29s	2.97± 0.17 13.72±1.70	1min43s

Table 3: Hall method *vs.* Clauset method. Number of subsamples is 200. To try different values of n_1, n_1/n is chosen to be some proportion of original size: $0.1, 0.2$ and 0.5. Also, n_1/n should be the proportion of power-law tail in whole data, which is known in simulation study but not known for applications. When $\alpha = 1.5, 2, 2.5, 3$, the corresponding proportions of power-law tail are $0.46, 0.24, 0.13$, 0.07, respectively.

$n_1 = 100$	α	γ
resample = 100. 80s	2.52±0.14	15.06±3.02
resample = 200. 153s	2.51 ±0.14	15.33±2.57
resample = 300. 230s	2.51± 0.13	15.27 ±2.51
resample = 400. 295s	2.51±0.14	15.12±2.40
resample = 500. 373s	2.51±0.14	15.15±2.71
resample = 600. 455s	2.51±0.13	15.19±2.48
resample = 700. 524s	2.51± 0.14	15.27±2.60
resample = 800. 592s	2.51±0.13	15.06±2.60
resample = 900. 657s	2.51±0.13	15.25±2.48

Table 4: Different number of subsamples in Hall method. When $\alpha = 2.5$, $n = 1000$, and $n_1 = 100$, an illustration which shows that the number of resamples does not make much difference for estimation.

- Use the formula provided in the Danielsson paper to calculate k:

$$k = \frac{k_1^2}{k_2}\left(\frac{(\log k_1)^2}{(2\log n_1 - \log k_1)^2}\right)^{(\log n_1 - \log k_1)/\log n_1}$$

Table 2 shows that Danielsson method takes much more time than Clauset method while providing no improvement for either bias or standard deviation. Therefore, even though Danielsson can achieve asymptotically minimized MSE, it does not perform well in smaller data size.

2.3.1.6 KS+rank The potential advantage of Danielsson method is that it avoids selection of the tuning parameter n_1 by examining all values of it automatically. However, this method turns out to give disappointing results despite much longer computation time. We therefor would like to find other methods to automatically select n_1. We propose two methods in this thesis: one we call KS+rank, and the other Clauset+Hall. Both combine ideas of the Clauset and Hall methods. We first introduce the KS+rank method.

The difference between Clauset method and Hall method is that the former searches values of γ directly and uses goodness of fit KS statistic as the standard of searching, while the latter method searches for rank of order statistic from where power-law tail starts, and uses the MSE standard instead. For estimating the MSE, the Hall method requires a subsample with a different size from original size; otherwise the bias part is usually estimated as 0 (Hall, 1990). If we do not use MSE as the standard, then estimated bias is not necessary, and we need not choose a subsample size smaller than original data size. Therefore, we keep Hall's idea of searching the rank and using bootstrap subsamples, but change the standard from MSE to KS statistic; then we can use subsample size same as original data size (which we call resamples rather than subsamples).

The procedure of KS+rank method is:

- Step 1: Suppose data set is $x_1, x_2, ..., x_n$. Determine then number of resamples, m. Simulate m resamples from given data set, each with size n.
- Step 2: Search all possible $k, k = 2, 3, ..., n$. For each k, and for each resample, treat the largest k values of the resample as coming from a power-law distribution. Use the MLE α and compute $KS_i (i = 1, 2, ..., m)$ for power-law tail part of each resample. Then compute the mean of KS_i across these resamples, denoted as KS^k.
- Step3: Choose k which minimizes KS^k.

In Step 1, we must determine number of resamples in the simulation. We tried several values, and (as with many bootstrap experiments) found that this number does not affect results much. Therefore in applications we can choose only $m = 10$ to make the process faster, which means the computation time of KS+rank is around 10 times of that of Clauset method (though in fact sometimes it takes 20 times of time for some extra time spent on calculations on the resamples). Table 2 indicates that KS+rank can decrease standard deviation well for both $\hat{\gamma}_{K+r}$ and $\hat{\alpha}_{K+r}$, however it will have larger bias for both of them too.

2.3.1.7 Clauset+Hall Another method to determine n_1 for Hall's method is based on the property we discovered from the previously mentioned simulation of Hall method: the best n_1/n is approximates the true proportion in the power-law tail. Although the true power-law proportion is not known without knowing γ, it could start iterations from an estimated γ to get an initial n_1, and use newly a estimated γ to estimate a new n_1, and iterate. In this thesis, considering computation time and unproved convergence of iterations, we iterate just once. We name this method as Clauset+Hall, because we used Clauset's method first to quickly get an estimate of γ, and used this estimated n_1 for Hall's method to get $\hat{\gamma}_{C+H}$ and $\hat{\alpha}_{C+H}$.

Table 2 indicates that Clauset+Hall can obtain nontrivial improvement for both bias and standard deviation for both $\hat{\gamma}_C$ and $\hat{\alpha}_C$ when $n = 500$ and 1000, with longer time consumption over Clauset. However, for $n = 100$, it performs less well than the Clauset's method. Though in Table 3 the result for $n = 100$ of Hall's method is still comparable with that of the Clauset's method, we need to remember that Clauset+Hall is based on $\hat{\gamma}_C$ from Clauset's method, which means that large bias and standard deviation of $\hat{\gamma}_C$ from Clauset's method will cause a large departure and variation in n_1 in Clauset+Hall leading to a large bias and standard deviation. When $n = 100$, estimates of Clauset's method have large bias and standard deviation, and therefore Clauset+Hall cannot have satisfactory results either.

2.3.1.8 Conclusions

- Clauset's method has comparatively large bias and standard deviation of estimates for some parameter settings. We proved that bias and standard deviation of $\hat{\gamma}_C$ will be proportional to the size of γ.

- The Jackknife decreases bias and standard deviation for both parameters, however the improvement is marginal, and it takes much more computational time.
- Hall's method, which can achieve minimized asymptotic MSE of $\hat{\alpha}_H$ theoretically, is better estimation than Clauset in both bias and standard deviation of both $\hat{\gamma}$ and $\hat{\alpha}$, if n_1 chosen appropriately. We observed that the best n_1 should be approximately the true amount of data in the power-law tail.
- Danielsson's method requires a much larger amount of time; despite that, its standard deviation is too large and very unstable, and performance of bias is also unstable.
- KS+rank has a smaller standard deviation but larger bias. It does not require n_1, and it is resistant to various choices of the number of resamples. Choosing a small number of resamples can make computation faster.
- Clauset+Hall decreases bias and standard deviation for both parameters visibly when $n = 500$ and $n = 1000$, without too much more time consumed. It does not require n_1. However, it has a worse result than Clauset when $n = 100$.
- Generally, searching rank of order statistic from which the tail becomes power-law distribution, such as Hall's method, KS+rank, and Clauset+Hall have smaller standard deviation of both $\hat{\gamma}_{C+r}$ and $\hat{\alpha}_{C+r}$ than the Clauset method.
- Searching rank method involves much more computing time than Clauset's method because it requires subsamples or resamples.

2.3.2 Type 4

2.3.2.1 Analysis and solutions There are three parameters to be estimated for Type 4: α, and the two boundaries of power-law behavior γ and ν. The procedure is similar to that for Type 3, which is to estimate ν first, then search for γ, and finally α using the estimated γ and ν.

Previously, Aban et.al (Aban et al., 2006) used $X_{(n)}$, which is MLE of ν, to estimate ν. They proposed to plot the estimated α against the searched rank of the order statistic k (largest k data is considered to follow power-law distribution), and choose k from which the estimated α becomes stable, which is empirically assessed. The k^{th} largest order statistic is the estimate of γ. They

pointed out that the choice of k affects the estimation of α. However, they suggested no automatic way to search for γ.

We tried to adapt methods in Type 3 to search for γ or k automatically. Since there is one more parameter ν to estimate, we use a more precise way to estimate parameters: in particular, we use a quasi-unbiased MLE for α, a modified upper limit for ν, and a stabilized KS-test to search for γ (Maschberger and Kroupa, 2009). (We tried such corrections for Type 3 too, but the results were quite similar to the one without correction. Therefore for Type 3 we used the original MSE and KS calculations.)

The quasi-unbiased MLE is:

$$\hat{\alpha}_{QUML} - 1 = \frac{n}{n-2}(\hat{\alpha}_{ML} - 1) \tag{2.10}$$

The modified upper limit is:

$$\hat{x}_{mmax} = X_{(n)}\left(1 + \frac{e^G - 1}{n}\right)^{\frac{1}{1-\hat{\alpha}_{QUML}}}, \quad \text{where} \quad G = (1 - \hat{\alpha}_{QUML}) \ln\left(\frac{X_{(n)}}{X_{(1)}}\right) \tag{2.11}$$

The stabilized KS-test statistic is:

$$SKS = \max_{1 \le i \le n} |S(\frac{i - 0.5}{n}) - S(P_{(i)})| \tag{2.12}$$

where P is the theoretical CDF, and $S(u) = 2S_0(0.5 + 0.5u) - 1$, $S_0(u) = \frac{2}{\pi}\arcsin(\sqrt{u})$. The stabilized KS-test deals with insensitivity of the KS-test in the tails.

Note that when calculating $\hat{\alpha}_{QUML}$ in Eq 2.10, we use $X_{(n)}$ as the estimate of ν, though finally estimate of ν is $\hat{x}_{mmax}$. We can not use $\hat{x}_{mmax}$ for calculation of $\hat{\alpha}_{QUML}$, because calculation of $\hat{x}_{mmax}$ in Eq 2.11 requires $\hat{\alpha}_{QUML}$ too.

We also adapt Clauset's method here: use $X_{(n)}$ as estimate of ν first, search through all possible values of γ, and for each searched γ fit a truncated power-law distribution to the tail by using the quasi-unbiased MLE from (2.10) and calculate the stabilized KS-test statistic from (2.12). $\hat{\gamma}_4$ is the one that minimizes the stabilized KS. Finally, we use (2.11) to get a modified estimate of ν denoted as $\hat{\nu}_4$.

Thus, the steps for the KS+rank adapted procedure are the following.

1. For a given data set, simulate several resamples, each with size as original data set size.

2. Search all possible k ($k = 3, 4..., n$). For each k, consider the largest k data from each resample following truncated power-law distribution, and get the quasi-unbiased MSE of α , modified upper limit ν, and the stabilized KS for the power-law tail of each resample. Calculate the mean KS for each k.
3. Choose k which minimizes KS.

We did not try Hall's method here, because there is no theoretical derivation for the relationship between the original data rank k and the subsample rank k_1 (recall that in Type 3, $k = nk_1/n_1$).

In simulations, we used the distribution with the form 2.13, which is just a truncated version of the simulation distribution for Type 3. Lemma 1 can still be applied for the truncated power-law distribution with the form (2.13). The interpretation is similar. For example, with α and n fixed, to compare estimates of two distributions: distribution A with $\gamma = 15, \nu = 50$, and distribution B with $\gamma = 150, \nu = 500$, then $b_B(\hat{\alpha}) = b_A(\hat{\alpha})$, $S_B(\hat{\alpha}) = S_A(\hat{\alpha})$, $b_B(\hat{\gamma}) = 10b_A(\hat{\gamma})$, $S_B(\hat{\gamma}) = 100S_A(\hat{\gamma})$, $b_B(\hat{\nu}) = 10b_A(\hat{\nu})$, $S_B(\hat{\nu}) = 100S_A(\hat{\nu})$, where b is bias, S is variance. Therefore, it is not necessary to try all combinations of γ and ν, since only the ratio between them matters. If there is a large ratio, the truncated distribution will be close to non-truncated distribution. Therefore we set $\gamma = 15$, and $\nu = 50$, whose ratio is not large, so that the result will be representative for typical truncated distributions. We also tried $\alpha = 1.5, 2, 2.5$, and $n = 100, 500, 1000$.

$$\begin{cases} p(x) = C(x/\gamma)^{-\alpha}, \text{ for } \gamma \le x \le \nu \\ p(x) = Ce^{-\alpha(x/\gamma - 1)}, \text{ for } x < \gamma \end{cases} \tag{2.13}$$

2.3.2.2 Results Table 5 shows that for estimation of ν, Clauset's method and KS+rank are similar. For the estimation of α and γ, Clauset is better than KS+rank in bias, while KS+rank is better than Clauset in standard deviation.

2.3.3 Density 6

2.3.3.1 Analysis and solutions There are three parameters to be estimated for type 6: γ, ν and α. Deluca et al. (Deluca and Corral, 2013) modified Clauset method to search pairs of γ and ν. However, they did not have a simulation study but simply applied it to real data whose values of parameters are unknown. In this thesis, we conducted a simulation study to evaluate their method.

$n = 100$	Clauset	time	KS+rank	time
$\alpha = 1.5$	1.57 $\pm$0.51		1.30$\pm$0.31	
$\nu = 50$	48.88$\pm$2.33		49.09 $\pm$2.35	
$\gamma = 15$	6.67 $\pm$3.88	1s	4.37$\pm$2.18	18s
$\alpha = 2$	1.94 $\pm$0.52		1.58$\pm$0.35	
$\nu = 50$	46.46$\pm$4.07		46.57$\pm$4.08	
$\gamma = 15$	6.16$\pm$2.95	1s	3.86$\pm$1.80	19s
$\alpha = 2.5$	2.25$\pm$0.61		1.78 $\pm$ 0.38	
$\nu = 50$	42.50$\pm$6.37		42.55 $\pm$6.34	
$\gamma = 15$	5.51$\pm$2.42	1s	3.37$\pm$1.46	18s
$n = 500$	Clauset	time	KS+rank	time
$\alpha = 1.5$	1.46 $\pm$ 0.20		1.37$\pm$ 0.17	
$\nu = 50$	49.66$\pm$0.65		49.66$\pm$ 0.65	
$\gamma = 15$	8.13$\pm$3.15	9s	6.57$\pm$ 2.41	1min54s
$\alpha = 2$	1.94 $\pm$0.31		1.79 $\pm$0.22	
$\nu = 50$	48.97$\pm$1.32		48.96 $\pm$1.32	
$\gamma = 15$	8.23$\pm$2.97	9s	6.45$\pm$ 1.88	1min49s
$\alpha = 2.5$	2.36$\pm$0.37		2.19$\pm$ 0.32	
$\nu = 50$	47.51$\pm$2.62		47.48$\pm$ 2.61	
$\gamma = 15$	7.95$\pm$ 2.83	9s	6.39 $\pm$1.84	1min54s
$n = 1000$	Clauset	time	KS+rank	time
$\alpha = 1.5$	1.49$\pm$ 0.24		1.43$\pm$0.14	
$\nu = 50$	49.82$\pm$ 0.27		49.82 $\pm$0.27	
$\gamma = 15$	9.85$\pm$4.08	21s	7.61 $\pm$1.96	4min33s
$\alpha = 2$	1.95 $\pm$0.23		1.90$\pm$0.24	
$\nu = 50$	49.45$\pm$0.59		49.44$\pm$ 0.59	
$\gamma = 15$	9.30 $\pm$3.25	22s	7.89 $\pm$2.25	4min17s
$\alpha = 2.5$	2.41 $\pm$0.26		2.30 $\pm$ 0.25	
$\nu = 50$	48.60$\pm$1.28		48.58$\pm$ 1.28	
$\gamma = 15$	9.25$\pm$2.53	23s	7.56 $\pm$ 1.74	4min5s

Table 5: Type 4, Clauset vs KS+rank. 100 data sets are simulated from (2.13). For KS+rank, number of resamples is $m = 10$.

Deluca's method is based on Clauset's method. The difference is that Deluca's method does not directly choose the pair of γ and ν which minimizes KS, rather, it selects a pool of candidates of pairs of γ and ν whose corresponding p-value from a goodness of fit test exceed a boundary, and then choose the pair which has the longest interval among candidates. The longest interval is chosen, for it might include many acceptable intervals. For this method, finding the p-value is very time consuming, because it cannot be calculated easily, so it requires bootstrap samples instead (which is also very time consuming when n is big).

We also tried Clauset's method and KS+rank here. Briefly speaking, Clauset's method searches the pair of γ and ν whose KS is smallest, and uses the MLE of α using data in this interval; KS+rank generates resamples from original data set, and determines k and s which are ranks of data, such that interval between k^{th} data and s^{th} largest data points following a power-law distribution. It chooses k and s whose KS is smallest, and uses the MLE of α for data in this interval.

We simulated data following distribution (2.14), which has smooth derivatives at the transition points; hence, it is quite challenging for estimation. Lemma 1 is applicable for distribution (2.14) too. Just as for Type 4, we need not try all combinations of γ and ν, since only the ratio between them matters. We set $\gamma = 15$. To decide the values of ν, α and β, we tried many combinations of them: $\alpha = 1.5, 2, 2.5, 3$, $\beta = 0.01, 0.1, 1, 10, 100$, $\nu = 50, 500$, to choose a combination of parameters which allows that the proportions of each part are not too small. Finally, we use $\gamma = 15, \nu = 50, \alpha = 1.5, \beta = 0.01$, in which the proportion for the components is: 59% for the head-impure part, 23% for the middle power-law part, and 18% for the tail-impure part.

$$\begin{cases} p(x) = Ce^{-\alpha(x/\gamma - 1)}, \text{ for } 0 < x < \gamma \\ p(x) = C(x/\gamma)^{-\alpha}, \text{ for } \gamma \leq x \leq \nu \\ p(x) = C(x/\gamma)^{-\alpha} e^{-0.1(x/\nu - 1)^2}, \text{ for } x > \nu. \end{cases} \tag{2.14}$$

2.3.3.2 Results Table 6 shows that Deluca's method has comparatively smallest standard deviation, however it has a rather intolerably large bias for all parameters. Clauset's method and the KS+rank method have larger standard deviations than Deluca's method, while having a critically smaller bias than Deluca's method.

	γ	ν	α	time
True value	15	50	1.5	
Deluca, n=100	1.00±0.66	203.56±53.00	1.08±0.11	7min39s
Clauset, n=100	5.53 ± 3.75	85.35±75.07	1.29±0.33	7min34s
ks+rank, n=100	4.46±2.24	132.15±65.36	1.322±0.20	1hr11min
Clauset, n = 1000	8.82±4.27	86.97±59.15	1.40±0.21	14hrs
ks+rank, n=1000	6.96±1.71	113.23 ±20.26	1.36±0.04	162hr40min

Table 6: Parameter estimation for Type 6. Simulation with the density format (2.14), with $\gamma = 15, \nu = 50, \alpha = 1.5, \beta = 0.01$ (β is only for simulation, not an estimating parameter). We simulate 100 data sets, each with sample size $n = 1000$.

Comparing the Clauset and KS+rank methods, we can see that when $n = 100$, Clauset is better in terms of bias of $\hat{\gamma}$ and $\hat{\nu}$, and KS+rank is better in bias of $\hat{\alpha}$ and in standard deviations of all parameters; when $n = 1000$, Clauset is better in bias of all parameters, and KS+rank is better in standard deviation of all parameters. All these programs consume a large amount of computation time. We did not do simulation study for Deluca's method for $n = 1000$, because the computational time is infeasible.

2.4 SUMMARY

- We studied properties of simulated data for Types 3, 4 and 6, which allows us to reduce the work for setting parameters in the simulation.
- We pointed out that bias of γ in Clauset's method for Type 3 data is nearly $\gamma/3$ for some cases, and when $\alpha = 1.5$, the standard deviation of γ is very large.
- We tried several existing methods and proposed two new methods to decrease bias and standard deviation of estimate of γ. Existing methods include Jackknife, Hall and Danielsson methods; newly proposed methods are KS+rank and Clauset+Hall.
- We adapted Clauset and KS+rank methods to Type 4 data to automatically search for γ.
- We adapted Clauset and KS+rank methods to Type 6, and compared it to Deluca's method.

3.0 MODEL SELECTION

3.1 INTRODUCTION

Methods of parameter estimation differ according to the different types of densities listed above. Not knowing the density type when estimating the parameters may lead to the use of an inappropriate estimation method (Burroughs and Tebbens, 2001). In deciding the density type, whether the density is truncated or not is a key feature to be considered.

If the truncated density is fitted with a non-truncated density or *vice versa*, the estimation of the exponent will not be correct, which will in turn lead to incorrect estimates of key quantities, such as the first and second moment of the density. In addition, there might be other concerns about the misuse of methods in application to different density types. For example, suppose that the data concern values related to risks, where high values represent high risk. Then, treating a non-truncated density as truncated will miss the high risk values, so we could be unprepared for the high risk; on the other hand, treating a truncated density as non-truncated could make the estimated risk higher than it actually is, increasing the cost of risk prevention, with too many false positives. Therefore, differentiating density types, especially between truncation and non-truncation is necessary.

Different model selection methods provide different sensitivity (true positive) and specificity (true negative) when deciding between non-truncated and truncated models. It will be helpful if researchers know the sensitivity and specificity for each method so they can weigh them and choose appropriate methods to decide the density type. In this chapter, we offer researchers this information about several model selection methods. These models include test-based methods (exceedance test and likelihood ratio test) and criterion based methods (AIC and BIC). For simplicity, we analyzed whether a density is truncated or not, only for the case of pure densities, that is, to

differentiate the density type between Types 1 and 2. We present the results of our simulation studies in tables below.

3.2 PREVIOUS WORK

There are several previous studies about deciding whether a power-law density is truncated or not. One approach is to see if there is a drop-off at the tail of the log-log survival curve (Burroughs and Tebbens, 2001), as seen in Figure 3. This is derived from a property that log-log plot of survival curve is straight for the non-truncated density with the form (2.1), because

$$\log\left(S(x)\right) = (\alpha - 1)\log\gamma - (\alpha - 1)\log\left(x\right) \tag{3.1}$$

while log-log plot for the truncated density with the form Eq 2.2 has a drop-off, because

$$\log\left(S(x)\right) = \log\left(\frac{1}{\gamma^{1-\alpha} - \nu^{1-\alpha}}\right) + \log\left(x^{1-\alpha} - \nu^{1-\alpha}\right) \tag{3.2}$$

To do this, one would fit data by both non-truncated and truncated power-law distributions to see which curve fits better, and then observe whether there is a drop-off at the tail of the better fit curve. However, sometimes this procedure is not reliable. Even if a given data set follows a non-truncated power-law density, the better-fit model must be truncated because it has one more parameter, and in practice data sets are (of course) finite. When observing whether there is a drop-off of the fitted truncated curve, (3.1) and (3.2) show that as long as $x_{(n)}$ is not too large, the fitted truncated density will have a drop off, even though the original data is non-truncated. Therefore, people tend to judge some non-truncated cases as truncated, as seen from Figure 3. On the other hand, if distribution of data is truncated, with large $x_{(n)}$, the drop-off often seems too trivial to consider the distribution as truncated.

Aban, et al. (Aban et al., 2006) proposed a test-based model selection method: they used exceedance test. Consider the following hypothesis testing problem:

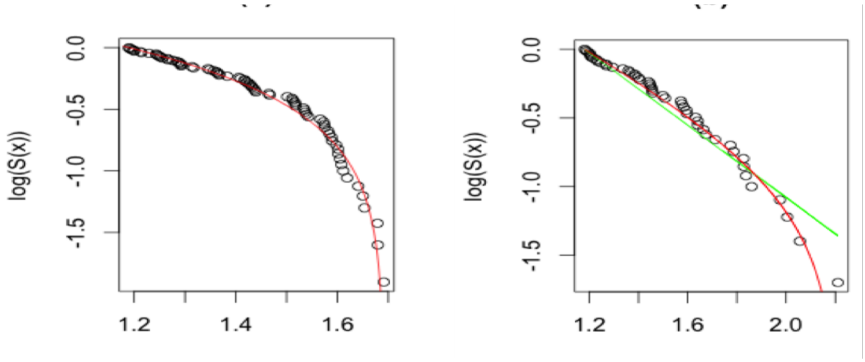

Figure 3: (a) is the log-log scale survival plot of a data sample from a truncated power-law distribution with form (2.2), where $\alpha = 1.5, \gamma = 15, \nu = 50$, sample size is 60. The red is fitted by a truncated power law. (b) is the log-log scale survival plot of a data sample from a non-truncated power law with form (2.1), where $\alpha = 1.5, \gamma = 15$, and sample size 50. The green line is the fitted by non-truncated and the red by truncated power law. Though data are non-truncated, because of the small value of $x_{(n)}$, there is still a drop-off in the tail and a truncated distribution is a better fit.

$$\mathrm{H}_0 : \text{ power-law distribution with } \nu = \infty \text{ (non-truncated power-law distribution)}$$

vs

$$\mathrm{H}_a : \text{ power-law distribution with } \nu < \infty \text{ (truncated power-law distribution)}$$

The significance level q for this test is the probability of deciding the power-law distribution as truncated when it is actually non-truncated. Here, we use the language of screening or classification: we treat non-truncated as "positive" and truncated as "negative" for convenience. The sensitivity (true positive) is $1 - q$, and the specificity (true negative) is the power of the test. For the exceedance test of (Aban et al., 2006), H_0 will be rejected if the largest order statistic is small. They did not analyze power in their paper, so that the specificity to detect a truncated power-law distribution is unknown.

Maschberger, et al. (Maschberger and Kroupa, 2009) considered several testing methods and conducted simulations to assess the powers of these methods. These testing methods include the following:

1. Empirical Distribution Function (EDF) methods based on distances, such as the KS statistic;
2. EDF methods based on the correlation measures, such as an R^2;
3. tests for exponentiality, because logarithm of power-law variate is exponential; and
4. tests specifically for truncation, including a likelihood ratio test and an exceedance test.

They used a simulation study to get empirical critical values during testing, and considered various parameter combinations of α, γ, ν and n. They concluded that among all of these tests, likelihood ratio test and exceedance test are most powerful; in some cases, however, such as when n is small, α is large, or ν is large, even likelihood ratio and exceedance tests can only achieve very low power (less than 0.1). In that paper, they confined the significance level of the test to be 0.05, which set the sensitivity of detecting non-truncated density as 95%. In real applications, a researcher may not pursue very high sensitivity of non-truncated density. Rather, depending on the purpose, they may balance sensitivity and specificity to choose a more appropriate significance level for the tests.

Deluca, et al. (Deluca and Corral, 2013) mentioned the possibility of using the AIC to do model selection. A truncated model has one more parameter than the non-truncation model, and AIC does adjust for the number of parameters in a model. However, they just mentioned it in a sentence without simulations or theoretical work to assess the performance of the AIC.

3.3 OUR SOLUTION

In this thesis, we consider two types of methods for model selection: test based methods and criterion based methods. For test based methods, suppose H_0: model is non-truncated *vs.* H_a: model is truncated. We use exceedance test and likelihood ratio test as testing methods, because they are the two most powerful tests in (Maschberger and Kroupa, 2009). For criterion based methods, we used AIC and BIC, because both of them are criteria to decide whether more parameters are necessary. Based on our theoretical and computational work, we created tables so that given parameters, researchers can use the tables to find sensitivity and specificity of each method and choose the most appropriate method.

There are several parameters involved in calculating sensitivity and specificity: α, γ, ν and n. Lemma 2 indicates that for the exceedance test, only the ratio between γ and ν matters for the

calculation for all four methods. Lemma 2 cannot be extended to the other three methods exactly, but by mimicking its proof of we have similar (approximate) conclusions for them. Therefore, we suggest only considering the ratio between γ and ν when using the table for all four methods.

Lemma 2. For the exceedance test, the power calculated for given q, α, n, γ and ν, is same as that calculated by q, α, n, $m\gamma$ and $m\nu$, where m is any positive constant, and q is significance level. More precisely, suppose that $X \sim \text{PL}(\alpha, \gamma, \nu)$, and $Y \sim \text{PL}(\alpha, m\gamma, m\nu)$, with γ and α known, ν unknown, where m is a positive constant. To conduct level-q hypothesis testing $\text{H}_0 : \nu = \infty$ vs. $\text{H}_a : \nu < \infty$ for both X and Y. For a sample size of n and $\nu = \nu_0$, power of the test for X is same as that of the test for Y.

Proof: The rejection region for X is $X_{(n)} < C_X$, such that $P(X_{(n)} < C_X | X \sim \text{PL}(\alpha, \gamma)) = q$; rejection region for Y is $Y_{(n)} < C_Y$, such that $P(Y_{(n)} < C_Y | Y \sim \text{PL}(\alpha, m\gamma)) = q$.
If $X \sim \text{PL}(\alpha, \gamma)$, then $mX \sim \text{PL}(\alpha, m\gamma)$. It follows that

$$\begin{aligned} P(X_{(n)} < C_X | X \sim \text{PL}(\alpha, \gamma)) &= P(mX_{(n)} < mC_X | mX \sim \text{PL}(\alpha, m\gamma)) \\ &= P((mX)_{(n)} < mC_X | mX \sim \text{PL}(\alpha, m\gamma)) \end{aligned}$$

It is easy to see that Y and mX have with same distribution $\text{PL}(\alpha, m\gamma)$, therefore $C_Y = mC_X$. Next, the power of the test for X is $\text{Power}_X = P(X_{(n)} < C_X | X \sim \text{PL}(\alpha, \gamma, \nu_0))$. If $X \sim \text{PL}(\alpha, \gamma, \nu)$, then $mX \sim \text{PL}(\alpha, m\gamma, m\nu)$. Thus it follows that

$$\begin{aligned} \text{Power}_X &= P(mX_{(n)} < mC_X | mX \sim \text{PL}(\alpha, m\gamma, m\nu_0)) \\ &= P((mX)_{(n)} < mC_Y | mX \sim \text{PL}(\alpha, m\gamma, m\nu_0)) \\ &= P(Y_{(n)} < C_Y | Y \sim \text{PL}(\alpha, m\gamma, m\nu_0)) \\ &= \text{Power}_Y \end{aligned} \tag{3.3}$$

We use several significance levels in the simulation: 0.01, 0.05, 0.1 and 0.15. We also tried different parameter combinations: $\alpha = 1.5, 2, 2.5, 3$, $\gamma = 15$, $\nu = 50, 150, 500$, $n = 30, 50, 100, 1000$. For each parameter combination, we simulated data to estimate the empirical critical value, and simulated data again to estimate power.

For the test based methods, the simulation procedure is the following.

1. Simulate 1000 non-truncated data sets with given values of α, n, and γ. Get the largest order statistic $x_{(n)}$ (or likelihood ratio) for each data set.
2. Find the empirical critical value such that the proportion of $x_{(n)}$ (or likelihood ratio) less than or equal to it (for LR, greater than) is the significance level q.
3. Simulate 1000 truncated data sets with the same given values of α, n, γ, as well as another parameter ν. Get largest order statistic again $x_{(n)}$ (or LR) for each data set.
4. The proportion of $x_{(n)}$ (or LR) less than or equal to (for LR, greater than) that empirical critical value is the power.

For criterion based methods, simulations are based on different parameter combinations. We tried criteria such as AIC and BIC respectively (Clarke et al., 2009). In this context, sensitivity is the proportion of times that a non-truncated power-law distribution is detected as non-truncated, and specificity is the proportion of times that a truncated power-law distribution is detected as truncated. To calculate sensitivity, the procedure is:

1. Simulate 1000 non-truncated data sets with given values of α, n, and γ.
2. For each eata set, apply AIC or BIC criteria to determine whether it is truncated or not.
3. The sensitivity is the proportion of determined non-truncated data sets.

To calculate specificity, the procedure is the following.

1. Simulate 1000 truncated data sets with given values of α, n, γ and ν.
2. For each data set, apply AIC or BIC criteria to determine whether it is truncated or not.
3. Specificity is the proportion of determined truncated data sets.

Table 7 and Table 8 provide the power under each significance level and each parameter combination, for both the exceedance test and likelihood ratio test. To use the tables, researchers may check the value of q and see the power under the most appropriate parameter combination. Note that when considering γ and ν, only the ratio between them matters, especially for exceedance test. After weighing the balance of type 1 error and power, researchers can choose an appropriate level q and testing method to do the test-based model selection.

Table 9 and Table 10 provide sensitivity and specificity under each parameter combination per criterion. To use these tables, researchers may choose the most appropriate parameter combination.

q = 0.01		$\alpha = 1.5$		$\alpha = 2$		$\alpha = 2.5$		$\alpha = 3$		q = 0.05		$\alpha = 1.5$		$\alpha = 2$		$\alpha = 2.5$		$\alpha = 3$	
n	ν	exc	LR	exc	LR	exc	LR	exc	LR	n	ν	exc	LR	exc	LR	exc	LR	exc	LR
30	50	100.0	86.3	100.0	53.5	100.0	18.6	15.5	13.5	30	50	100.0	97.7	100.0	82.2	100.0	55.0	72.7	26.8
30	150	100.0	69.1	25.4	12.1	1.8	2.1	0.6	0.8	30	150	100.0	89.5	100.0	38.9	14.1	14.8	6.9	6.5
30	500	100.0	28.4	2.2	3.4	1.3	1.2	0.3	0.5	30	500	100.0	65.9	9.4	15.3	5.4	6.3	4.3	4.6
50	50	100.0	99.3	100.0	96.2	100.0	73.0	100.0	33.3	50	50	100.0	100.0	100.0	99.7	100.0	92.8	100.0	69.0
50	150	100.0	91.7	100.0	34.1	3.2	5.2	1.8	1.1	50	150	100.0	99.4	100.0	75.8	23.1	21.4	7.6	7.0
50	500	100.0	79.6	6.8	2.4	1.7	1.8	1.0	0.7	50	500	100.0	93.7	21.8	15.2	8.3	7.0	4.3	4.2
100	50	100.0	100.0	100.0	100.0	100.0	99.8	100.0	80.8	100	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.2
100	150	100.0	100.0	100.0	88.5	16.4	24.7	2.6	2.5	100	150	100.0	100.0	100.0	99.8	100.0	56.3	13.9	10.3
100	500	100.0	99.9	12.6	14.9	1.8	1.3	1.0	1.1	100	500	100.0	99.9	94.0	54.3	8.1	8.5	5.7	5.2
1000	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1000	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1000	150	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1000	150	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1000	500	100.0	100.0	100.0	100.0	100.0	68.3	2.3	2.4	1000	500	100.0	100.0	100.0	100.0	100.0	99.8	15.4	12.1

Table 7: Table of power for test based methods, for $q = 0.01, 0.05$

q = 0.1		$\alpha = 1.5$		$\alpha = 2$		$\alpha = 2.5$		$\alpha = 3$		q = 0.15		$\alpha = 1.5$		$\alpha = 2$		$\alpha = 2.5$		$\alpha = 3$	
n	ν	exc	LR	exc	LR	exc	LR	exc	LR	n	ν	exc	LR	exc	LR	exc	LR	exc	LR
30	50	100.0	98.8	100.0	93.3	100.0	78.6	100.0	58.2	30	50	100.0	99.4	100.0	97.5	100.0	88.6	100.0	70.6
30	150	100.0	96.2	100.0	56.8	31.7	26.6	13.5	13.4	30	150	100.0	97.0	100.0	70.0	34.6	38.5	22.3	23.3
30	500	100.0	79.8	26.3	24.3	10.6	10.6	11.3	12.0	30	500	100.0	87.7	37.5	37.1	19.9	16.7	16.5	14.5
50	50	100.0	100.0	100.0	99.9	100.0	96.5	100.0	86.2	50	50	100.0	100.0	100.0	100.0	100.0	98.9	100.0	93.8
50	150	100.0	100.0	100.0	89.8	47.7	45.3	14.3	14.5	50	150	100.0	100.0	100.0	93.2	82.1	55.9	25.1	23.2
50	500	100.0	96.5	4.5	38.7	15.2	17.2	10.4	11.6	50	500	100.0	99.8	66.2	49.7	20.1	21.2	18.1	15.1
100	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
100	150	100.0	100.0	100.0	100.0	100.0	82.1	26.6	28.5	100	150	100.0	100.0	100.0	99.9	100.0	88.5	42.8	41.1
21.7	500	100.0	100.0	100.0	77.6	14.2	15.4	11.2	11.2	100	500	100.0	100.0	100.0	86.5	21.7	28.3	14.0	16.9
1000	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1000	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1000	150	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1000	150	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1000	500	100.0	100.0	100.0	100.0	100.0	100.0	25.1	22.2	1000	500	100.0	100.0	100.0	100.0	100.0	100.0	37.4	42.2

Table 8: Table of power for test based methods, for $q = 0.1, 0.15$

Sensitivity		$\alpha = 1.5$		$\alpha = 2$		$\alpha = 2.5$		$\alpha = 3$	
n	ν	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
30	50	59.0	78.2	58.8	79.8	61.2	82.1	63.1	81.5
30	150	63.4	82.1	61.7	80.5	61.9	82.3	61.7	81.5
30	500	61.6	80.5	60.1	79.3	58.6	79.8	64.0	83.8
50	50	60.5	84.4	60.5	85.6	63.4	86.8	60.7	85.2
50	150	60.5	85.7	61.1	84.6	59.7	83.4	60.7	85.5
50	500	61.4	84.8	60.7	84.4	63.9	87.5	59.4	84.2
100	50	61.6	88.6	64.3	90.6	61.5	89.6	60.6	88.7
100	150	64.7	90.3	60.0	90.0	61.7	89.3	61.1	89.2
100	500	61.1	88.9	60.7	89.8	61.6	88.5	62.6	89.6
1000	50	63.1	96.6	65.0	96.9	63.8	96.8	62.3	96.0
1000	150	63.7	97.0	63.2	96.2	63.7	96.5	60.9	95.9
1000	500	65.9	97.4	62.9	96.3	60.8	96.4	64.1	97.7

Table 9: Table of sensitivity for criteria based methods

Specificity		$\alpha = 1.5$		$\alpha = 2$		$\alpha = 2.5$		$\alpha = 3$	
n	ν	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
30	50	100.0	99.9	99.8	97.8	98.9	92.1	94.9	76.0
30	150	99.8	98.2	96.1	80.5	71.5	39.7	48.5	25.0
30	500	98.9	94.4	72.5	42.3	46.2	24.3	41.0	20.2
50	50	100.0	100.0	100.0	100.0	100.0	98.8	99.2	92.8
50	150	100.0	100.0	99.7	93.5	90.3	53.2	61.3	24.4
50	500	100.0	99.6	85.9	50.4	49.8	21.0	40.4	17.2
100	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
100	150	100.0	100.0	100.0	100.0	99.7	80.3	81.0	25.4
100	500	100.0	100.0	99.7	79.2	62.0	16.8	41.8	12.8
1000	50	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1000	150	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1000	500	100.0	100.0	100.0	100.0	100.0	99.5	86.9	8.1

Table 10: Table of specificity for criteria based methods

They check sensitivity and specificity for each criterion and decide which criterion to use for the model selection.

We conclude by noticing the following features in these tables.

1. Table 7 and Table 8 show that larger q, larger n, smaller ν and larger α can increase power, for both the exceedance test and likelihood ratio test.
2. Table 9 and Table 10 show that larger q, larger n and smaller ν can increase specificity while decreasing sensitivity. Sensitivity and specificity of both AIC and BIC, especially AIC, are fairly stable with changes in α, as the power is in test based methods.
3. Table 9 and Table 10 show that BIC has higher sensitivity than AIC, while AIC has a higher specificity than BIC. This is a reasonable finding, because BIC prefers a more parsimonious model than AIC.

3.4 SUMMARY

In this part, we provided four model selection methods. We created tables so researchers can select an appropriate method after weighing the sensitivity and specificity. We observed some trends in the tables. We also proved Lemma 2 for the exceedance test.

4.0 RESILIENCE

4.1 INTRODUCTION

Power-law distributions have many applications, one of which is that the degree distribution of many networks follow power-law distributions. The degree of a vertex of a network is the number of connections it has to other vertices. The degree distribution is the proportion of these degrees distributed over the entire network. If the degree distribution of a network follows a power-law distribution, at least approximately, then the network is called a scale-free network because changing the scale does not change the exponent of its power-law distribution. Figure 3 shows an example of a power-law degree distribution plot. Many networks are found to be scale-free, such as the world wide web's (WWW's) internet links, certain biological networks, and some social networks (Barabási et al., 2009).

The degree distribution is a critical characteristic of networks which gives topological information and provides a basis for calculations of other important quantities. For example, in protein-protein interaction networks, investigating high-degree proteins as drug targets might provide a good approach for therapeutic mediation (Han et al., 2005). Such a strategy would have less of an impact if the true topology is exponential rather than power-law distribution. In epidemiology networks (Newman, 2002), the speed of the disease's spread is a function of the second moment of the degree distribution.

In many applications, networks receive attacks causing them to lose vertices and edges. For example, in the epidemiology network, people who get immunized will be removed from the network; in computer networks, hackers attack the network to make some computers malfunction or lose connection with other computers. There are many attack strategies for how to remove vertices or edges. In this thesis, we focus on the removal strategies for vertices, not for edges. We

assume that if the vertices are removed by a strategy, their edges will be removed together with them. In the current literature about network resilience (Cohen et al., 2000) (Holme et al., 2002), the typical removal strategies are random removal or high-degree removal, which are to remove vertices randomly or by the order of their degrees. In this thesis, we also explored an additional generalized removal strategy that we define below.

When a network loses vertices due to an attack, the degree distribution may change and, consequently, affect the functioning of the network. We study both theoretically and empirically how the degree distribution changes after vertices are removed, which is called assessing the resilience of the degree distribution after a network is attacked (the meaning of "resilience" will be explained further in next section). Through theoretical studies, we show below that the power-law degree distribution does not hold after only one vertex is removed from a power-law degree distribution, regardless of the removal strategy. However, by a simulation study, using KS test to determine whether degree distribution is power-law distribution will test it as still being power-law distribution even after larger proportions of vertices are removed, which indicates a discrepancy between theoretical and experimental work. Note, however, that the experimental work gives an inaccurate but tolerable conclusion: it is often appropriate to declare a degree distribution as following power-law distribution even if there is a small departure, because properties of power-law distribution are very useful, and may well be robust to small perturbations from the power law. We can imagine that in real studies, if researchers conclude (say, using the KS-test) that the degree distribution is still a power-law distribution after an attack, which is inaccurate but tolerable, they might be interested in how parameters will change for the power-law distribution. Therefore, we also conducted an numerical experiments to study the change of parameters when power-law distribution is tested to be still valid after attacks.

4.2 RELATED WORK

There are several studies about the resilience of a network different from our work. These studies are more about resilience of the connectivity of a network, so the measurements they use are tailored for that purpose. Such measurements include the critical fraction of nodes that need to

be removed before the network disintegrates (Cohen et al., 2000, 2001), size of the largest connected subgraph, or average inverse geodesic distance (the number of edges in the shortest path connecting them) (Holme et al., 2002). They consider a network is resilient if after removing a considerable number of vertices, the critical fraction of nodes for disintegration remains high, or the size of the largest subgraph remains large, or the average inverse geodesic distance remains large. They concluded that power-law networks are resilient to random removals: in particular, even when around 80% vertices are removed, the network still remains connected (Cohen et al., 2000). However, power-law networks are not resilient to the high-degree removal: in fact, the network becomes disconnected with high probability when only 5% vertices are removed (Cohen et al., 2001; Callaway et al., 2000).

Below, we use the definition of resilience of the connectivity of a network to define the resilience of the degree distribution of a network: the degree distribution is resilient if it does not change much after removing a considerable number of vertices from the network. We also adopt their simulation procedures to simulate attacks by removing an increasing proportion of vertices and observe the change in the degree distribution.

Next, there are studies exploring whether the degree distribution of the sampled network is the same as that of the whole network (Kolaczyk, 2009) (Stumpf et al., 2005). Sampling a network is defined as taking a sample of vertices and include all edges between them. There are sampling strategies for either sampling with replacement or without replacement. For those sampling strategies with replacement, there are various sampling probabilities, such as random sampling and degree dependent sampling. For those sampling strategies without replacement, there are various traversal techniques, such as breath/depth-first search and snowball sampling, which are procedures where newly selected nodes depend on other selected nodes. Our removal approach is different from sampling. We remove vertices without replacement with various removal probabilities. It is hard to find a relationship between the sampling strategies and removal strategies directly. There is no need to explore the relationship between sampling and removal probabilities, because one is with replacement and the other is without replacement. It is not easy either to find a relationship between their procedures both for without replacement, because the sampling strategies use connections with selected nodes but removal strategies do not use such information. Therefore, conclusions on the degree distribution from sampling research cannot be applied to the

research of resilience after removal. One future possible research direction might be designing removal strategies in the same style as sampling strategies. In that way, it might be possible to find a relationship between these two types of research.

4.3 SIMULATIONS AND ANALYTICAL DERIVATIONS

In this Section, we present two experiments as well as a theoretical derivation to explore the change of the power-law degree distribution after the network is attacked. The first experiment evaluates whether the degree distribution type will change using simulation and a hypothesis test. We then do theoretical derivations to deduce the exact form of the expected degree distribution after attacks. The second experiment explores how the parameters of the degree distribution change when a hypothesis test indicates that the distribution type remains a power-law after attacks. Therefore, by looking at both the distribution type and distribution parameters, we have a full range of exploration of the change in the degree distribution due to attacks, or removal of vertices.

4.3.1 Experiment 1

For the first experiment, we try three attack strategies: random removal, high degree removal, and normal curve removal. Removal strategies are described by their removal probability densities, which models the probability that a vertex of a certain degree is removed. For random removal, the removal density is a uniform curve (Figure 4a), which means that the chance to remove a vertex is the same for all vertices. For high degree removal, the density is piecewise constant, with vertices of high degree having a high chance to be removed: in Figure 4b a vertex with the degree above (below) some threshold will definitely (not) be removed. For normal curve removal, vertices medium-sized degree are removed with the highest probabilities.

Random and high degree removal are commonly studied removal strategies. We propose a new strategy, the normal curve removal, because we believe that it should be a realistic strategy. It combines two approaches to attacks. One is that the larger the degree of a vertex is, the easier this vertex can be spotted and removed. For example, hackers seldom pay attention to individual

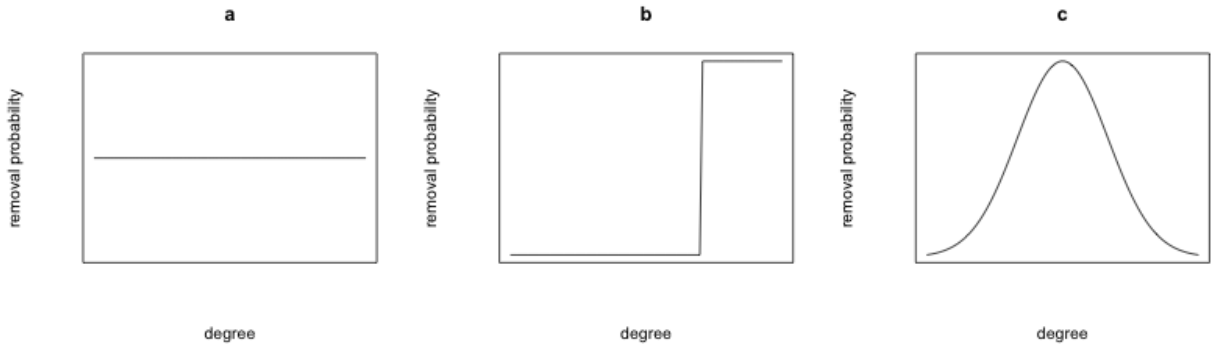

Figure 4: Removal probability models for (a) random removal; (b) high-degree removal; (c) normal curve removal.

accounts, but attack some hubs of computers if they want to destroy a computer network. The other is that the larger the degree of vertex is, the harder it is for the attacker to remove the vertex. For example, hackers may not be able to attack any important hubs because they have very good security. Thus, they may only attack small-size accounts. Combining these two perspectives, we propose the normal curve removal, which models the situation in which attackers pay more and more attention to remove a vertex when the degree of the vertex increases, but it gets harder to remove such a vertex also. In the simulation, the normal curve is in the domain $(1, \text{degree}_{\text{max}})$, with the same shape as the standard normal curve in the domain $(-3, 3)$.

We tried three networks. The first two are simulated networks and the third is a real network. Network 1 is Babarasi-Albert network, which is considered a typical example for power-law degree distribution with $\alpha = 3$. Network 2 is a simulated network by configuration method, that is, generating a network with a given expected degree distribution. Network 3 is a protein-protein interaction network, whose degree distribution is considered to follow power-law distribution (Han et al., 2005). There are 10000, 2000, and 1870 vertices in the B-A, configuration, and the protein networks, respectively.

We simulate attacks by removing an increasing proportion of vertices to mimic an attack that is increasingly severe. For random removal and normal curve removal strategies, we remove 5% to 95% of the vertices, in steps of 5%. We do this because power-law distribution prevails even when

a large proportion of vertices are removed. For the high-degree removal, however, the power-law distribution collapses very quickly when only small proportion of vertices are removed. To better see the details, we remove 1% to 19% of the vertices, in steps of 1%.

The simulation procedure is as follows: determine what vertices will be removed according to their removal probabilities for each removal strategy. Remove an increasing proportion of vertices. After each removal, determine the degree distribution of the remaining network by fitting the degree distribution and using KS test to test the goodness of fit. We fit not only a power-law distribution, but also five other models: a power-law distribution with a cutoff, Poisson, exponential, stretched exponential (Weibull) and lognormal distributions, because these are commonly used distributions for fitting degree distributions. We consider a large p-value from the goodness of fit test as a good fit of the model.

The results are shown in Figure 5, from which we draw the following conclusions:

1. With random and curve removals, for all three networks, the degree distribution type remains power-law distribution even when over 90% of the vertices are removed.
2. With high-degree removals for all three networks, a power-law distribution does not hold even after only about 5% of the high-degree vertices are removed.
3. With high-degree removal, the networks lose long tails, and an exponential distribution is a better fit.
4. For network 3, when more than 50% of vertices are removed by random removal or normal curve removal, the exponential distribution becomes the best model.
5. This is consistent with the resilience research of the network for network connectivity; that is, the connectivity is quite resilient to random removal but very fragile to high-degree removal.

4.3.2 Analytic Development

4.3.2.1 Introduction. Experiment 1 illustrates the changes in degree distributions by simulation. We also derived the exact form of the expected degree distribution after removing vertices. We derived formulas for all three removal strategies, and for each one, we developed the formula starting from the removal of one vertex and extending to any number of vertices. No assumption is required for random removal; however certain assumptions are needed for normal curve and high

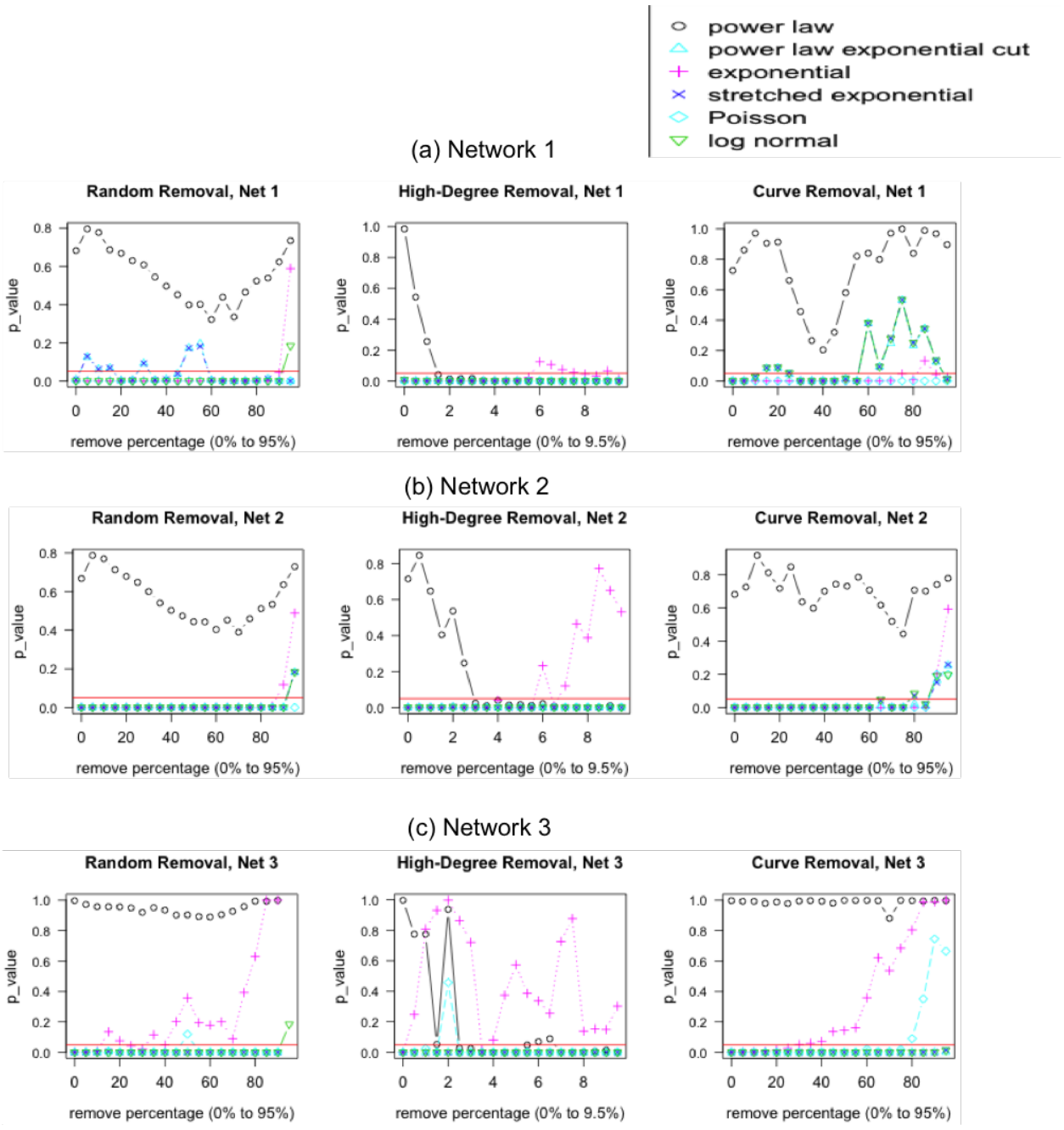

Figure 5: Fit of degree distribution by six models. There are three networks and three removal strategies. For random and normal curve removal, the procedure is to remove vertices from 5% to 95%, by 5% each time. For high degree removal, the procedure is to remove vertices from 1% to 10%, by 1% each time. The y-axis is the p-value of the goodness of fit test. The red horizontal line is the 0.05 threshold. p-values of different models is depicted by different colors and marks.

degree removal. Although the degree distribution may not reject the power-law hypothesis after attacks in some cases, the mathematical form of the degree distribution is not power-law distribution after even one vertex removed, regardless of removal strategy.

For any vertex v, let $d(v)$ denote the degree of v, and let $\{p_k\}$ denote the degree distribution of a network with n vertices, where $p_k = P(d(v) = k)$, the proportion of vertices with degree k among n vertices, for $k = 1, 2, ..., n-1$. When the degree distribution is a power-law distribution, $p_k = Ck^{-\alpha}$ for some α.

Denote the expected degree distribution of the new network by $\{q_k\}$. The derivation starts from removing only one vertex randomly, and then moves to removing r vertices randomly from the n vertices, where $r = 1, 2, 3, ..., n-1$. Usually people do not consider the situation of 0 degree, however in our derivation we included 0 degree, just for providing more information. It is easy to remove the situation of 0 degree by multiplying a normalization constant for expected degree of other degrees, which will not change the fact that whether expected degree distribution is power-law distribution or not.

$q_k = E(\text{Prob}(d(v) = k)|\text{one vertex removed})$, which is the expected proportion of vertices with degree k among $n-1$ vertices, $k = 1, 2, ..., n-2$. The expected value q_k is a weighted sum of all possible values. Each possible value is a proportion of vertices of degree k in the remaining network. Suppose the information known about the original network is the number of vertices and the degree distribution, therefore any network with the same information should be taken into consideration. We denote each possible value of q_k as $q_k^{T_M,M,W}$, where $W = 1, 2, 3, ...$ is the index of networks which have same information as the original network, $M = 1, 2, ..., n-1$ is the degree of the removed vertex, $T_M = 1, 2, ..., M$ is the index of the vertex removed among all removed vertices whose degrees are M. For example, $q_k^{3,5,2}$ is the proportion of vertices with degree k when the network is the second network, the removed vertex's degree is 5, and the removed vertex is the third one in all removed vertices with degree 5 in the second network (here orders of index of the network and vertices can be created by any means).

Note that $q_k = \text{E}(\text{E}(q_k^{T_M,M,W}|W))$. We derive $\text{E}(q_k^{T_M,M,w}|W = w)$ first, where w is any given network. We find that w does not appear in $\text{E}(q_k^{T_M,M,w}|W = w)$, therefore q_k is simply $E(q_k^{T_M,M,w})$.

Our notation used in this derivation are the following.

- V is the set of vertices, $|V| = n$. V^w is set of vertices for network w.

- v and v' are vertices in V. $v^{t_m,m,w}$ is the removed vertex from network w with degree m and index t_m.
- $d(v)$ is the degree of vertex v, that is, number of links that v has.
- $l(v', v) = 1$ means that v' and v are connected (0 means unconnected). Note that $l(v, v) = 1$.
- $\pi(v)$ is the removal probability of vertex v.
- Networks considered in this thesis are simple networks: they are undirected, without multiple edges between any two vertices, or loops within any vertex.

4.3.2.2 Random removal: Remove one vertex. Suppose that the degree distribution is $p_k = Ck^{-\alpha}$, for $k = 1, 2, ..., n-1$. For random removal, the removal probability for any vertex is $1/n$. After removing a particular vertex $v^{t_m,m,w}$ from among the n vertices, $q_k^{t_m,m,w}$ is the proportion of vertices with degree k among $n-1$ vertices of the remaining network. Vertices with degree k after the removal include two types of vertices: first, there are vertices with degree $k+1$ before and with degree k after the removal; the rest are vertices with degree k before and remaining degree k after, as seen in line 1 of (4.1).

The summations in line 3 of (4.1) are not easy to evaluate. Take the first item for example: it is not easy to directly calculate the summation, across all removal situations (that is, all values of M and T_M), of how many vertices are with degree $k+1$ before and with degree k after removal. We therefore found another way to calculate it: since the removal probability is same for all vertices, it is equivalent to calculating the summation (see the first item in line 1 of (4.3)) across all vertices with degree $k+1$ in network w of how many removals can remove one degree from them. The latter is easy to get (see the first item in line 2 of (4.3)): for each vertex with degree $k+1$ before (there are totally np_{k+1} such vertices) and k after, the number of the removals which removes one degree from this vertex is $k+1$, because each of its linked vertices can be removed. Similarly, the second item in line 1 of (4.3) can be derived too. For each vertex with degree k both before and after the removal, the number of removal situations are $n-(k+1)$, because each vertex not linked with this vertex can be removed, as well as itself.

Now, (4.3) shows that $\mathrm{E}(q_k^{T_M,M,w}|W = w)$ is independent of w, therefore

$$q_k = E(E(q_k^{T_M,M,W})|W) = E(q_k^{T_M,M,w})$$

for any w. (4.3) is the result for expected degree distribution q_k, as shown in (4.3). From (4.3) we conclude that even after one vertex is removed, the degree distribution of the network is no longer a power-law distribution. Next, we turn to the evaluation of the conditional expectation above.

$$E(q_k^{T_M,M,w}|W=w) =$$
$$\sum_{m=1}^{n-1}\sum_{t_m}\pi(v^{t_m,m,w})\frac{\sum_{v^{t_m,m,w}v'\in V^w}(\mathbb{1}_{d(v')=k+1,l(v',v^{t_m,m,w})=1}+\mathbb{1}_{d(v')=k,l(v',v^{t_m,m,w})=0})}{n-1}$$

$$= \sum_{v\in V^w}\pi(v)\frac{\sum_{v'\in V^w}(\mathbb{1}_{d(v\prime)=k+1,l(v',v)=1}+\mathbb{1}_{d(v')=k,l(v',v)=0})}{n-1} \tag{4.1}$$
$$= \frac{1}{n(n-1)}(\sum_{v\in V^w}\sum_{v'\in V^w}\mathbb{1}_{d(v')=k+1,l(v',v)=1}+\sum_{v\in V^w}\sum_{v'\in V^w}\mathbb{1}_{d(v')=k,l(v',v)=0})$$

$$\begin{aligned}
(4.1) &= \frac{1}{n(n-1)}(\sum_{v'\in V^w}\sum_{v\in V^w}\mathbb{1}_{d(v')=k+1,l(v',v)=1}+\sum_{v'\in V^w}\sum_{v\in V^w}\mathbb{1}_{d(v')=k,l(v',v)=0})\\
&= \frac{1}{n(n-1)}((k+1)np_{k+1}+(n-(k+1))np_k\\
&= \frac{k+1}{n-1}p_{k+1}+\frac{n-(k+1)}{n-1}p_k \qquad (4.2)\\
&= \frac{k+1}{n-1}C(k+1)^{-\alpha}+\frac{n-(k+1)}{n-1}Ck^{-\alpha}
\end{aligned}$$

Therefore,

$$q_k = \begin{cases} \frac{p_1}{n-1} & k=0 \\ \frac{(k+1)p_{k+1}+(n-k-1)p_k}{(n-1)} & k=1,2,...,n-2 \end{cases} \tag{4.3}$$

Remove two vertices.

The derivation of the expected degree distribution q_k after random removal of two vertices is quite similar to that of the removal of one vertex. It is still a weighted sum of all possible degree distribution values for degree k. The weight for each removal is $\binom{n}{2}^{-1}$, and the value of the degree distribution is the proportion of vertices with degree k among $n-2$ vertices. Vertices with degree k after removal include three types of vertices: vertices with degree $k+2$ before and with degree k after the removal, vertices with degree $k+1$ before and with degree k after, and vertices with degree k before and remaining degree k after. Again, we exchange the orders of the double summations. We denote two removed vertices as v_1 and v_2. It follows that

$$q_k = \sum_{v_1,v_2 \in V^w} \pi(v_1,v_2) \frac{\sum_{v_2 \in V^w} \big(\mathbb{1}_{d(v')=k+2, l(v',v_1)=1, l(v',v_2)=1} + \mathbb{1}_{d(v')=k+1, l(v',v_1)=1 \text{or} l(v\prime,v_2)=1}}{n-2}$$

$$\frac{+\mathbb{1}_{d(v')=k, l(v',v_1)=0, l(v',v_2)=0} \big)}{n-2}$$

$$= \frac{1}{\binom{n}{2}(n-2)} \Big(\sum_{v' \in V^w} \sum_{v_1,v_2 \in V^w} (\mathbb{1}_{d(v_2)=k+2, l(v',v)=1, l(v',v_2)=1} + \mathbb{1}_{d(v_2)=k+1, l(v',v)=1 \text{or} l(v',v_2)=1}$$

$$+ \mathbb{1}_{d(v')=k, l(v',v_1)=0, l(v',v_2)=0} \big)$$

$$= \frac{1}{\binom{n}{2}(n-2)} \left[\binom{k+2}{2} n p_{k+2} + \binom{k+1}{1} \binom{n-(k+2)}{1} n p_{k+1} + \binom{n-(k+1)}{2} n p_k \right] \quad (4.4)$$

$$q_k = \begin{cases} \frac{n}{\binom{n}{2}(n-2)} (p_2 + (n-2)p_1) & k = 0 \\ \frac{n}{\binom{n}{2}(n-2)} \Big(\binom{k+2}{2} p_{k+2} + \binom{k+1}{1} \binom{n-(k+2)}{1} p_{k+1} + \binom{n-(k+1)}{2} p_k \Big) & k = 1, ..., n-3 \end{cases} \quad (4.5)$$

Remove any number of vertices.

Suppose we randomly remove r vertices ($0 < r < n$). By the same logic we can derive the expected degree distribution: $q_k = \frac{1}{\binom{n}{r}} \frac{I_1 + I_2 + ... + I_{r+1}}{n-r}, k = 0, 1, ..., n-r-1$ where

$$I_{j+1} = \binom{n-(k+j+1)}{r-j} \binom{k+j}{j} n p_{k+j}, \text{ for } j = 0, 1, ..., r,$$

which is how many removals cause vertices with initial degree $k+j$ to lose j degrees to have degree k after the removal. Note that when $k = 0$, $I_1 = 0$.

Check the correctness of the derivation.

Here we check that $\sum_k q_k = 1, (k = 0, 1, 2, ..., n-r-1)$ is satisfied, when r vertieces are removed.

Check for one vertex removed.

$$\sum_{k=0}^{n-2} q_k = \frac{1}{n-1} p_1 + \sum_{k=1}^{n-2} \frac{k+1}{n-1} p_{k+1} + \frac{n-(k+1)}{n-1} p_k \quad (4.6)$$

The summation in Eq 4.6 is a linear combination of $p_k, k = 1, 2, ..., n-1$. When $k = 1$, the coefficient of p_k is $\frac{1}{n-1}+\frac{n-2}{n-1} = 1$; when $2 \leq k \leq n-2$, the coefficient of p_k is $\frac{n-(k+1)}{n-1}+\frac{k-1+1}{n-1} = 1$; when $k = n-1$, the coefficient of p_k is $\frac{n-2+1}{n-1} = 1$. Therefore,

$$\sum_{k=0}^{n-2} q_k = \sum_{k=1}^{n-1} p_k = 1$$

Check for two vertices removed.

$$\begin{aligned}\sum_{k=0}^{n-3} q_k &= \frac{n}{\binom{n}{2}(n-2)}(p_2 + (n-2)p_1) \\ &+ \sum_{k=1}^{n-3} \frac{n}{\binom{n}{2}(n-2)}(\binom{k+2}{2}p_{k+2} + \binom{k+1}{1}\binom{n-(k+2)}{1}p_{k+1} + \binom{n-(k+1)}{2}p_k)\end{aligned} \tag{4.7}$$

The summation in (4.7) is also a linear combination of $p_k, k = 1, 2, ..., n-1$. Similarly, when $k = 1$, the coefficient of p_k is $\frac{n}{\binom{n}{2}(n-2)}((n-2) + \binom{n-(1+1)}{2}) = 1$; when $k = 2$, the coefficient of p_k is $\frac{n}{\binom{n}{2}(n-2)}(1 + \binom{n-(2+1)}{2} + \binom{1+1}{1}\binom{n-(1+2)}{1}) = 1$; when $3 \leq k \leq n-3$, the coefficient of p_k is $\frac{n}{\binom{n}{2}(n-2)}(\binom{n-(k+1)}{2} + \binom{k-1+1}{1}\binom{n-(k-1+2)}{1} + \binom{k-2+2}{2}) = 1$; when $k = n-2$, the coefficient of p_k is $\frac{n}{\binom{n}{2}(n-2)}(\binom{n-3+1}{1}\binom{n-(n-3+2)}{1} + \binom{n-4+2}{2}) = 1$; when $k = n-1$, the coefficient of p_k is $\frac{n}{\binom{n}{2}(n-2)}\binom{n-3+2}{2} = 1$. Therefore,

$$\sum_{k=0}^{n-2} q_k = \sum_{k=1}^{n-1} p_k = 1$$

Check for any number of vertices removed.

Using a calculation similar to the one above, we check that coefficients of $p_k, k = 1, 2, ..., n-1$ are all 1 so that $\sum_{k=0}^{n-r} q_k = \sum_{k=1}^{n-1} p_k = 1$. We omit writing the details here.

Comparison between the derived expected degree distribution and power-law distribution. Not only did we derive the formulas above, we also demonstrated more intuitively the degree distributions after removal of vertices by plotting the derived expected degree distributions and the original power-law degree distribution together. Recall that the mathematical form indicates that

power-law distribution is destroyed even only one vertex is randomly removed, however, in the simulation study the power-law distribution is quite resilient even after 90% vertices are removed, by using the KS-test of goodness of fit. It is possible that the expected degree distribution seems similar to power-law distribution such that KS is not able to detect the difference even by significance level 0.05. It is also possible that the expected degree distribution is not similar to the original degree distribution even though it passes a power-law test; rather, it might be fit well by power-law distribution with different parameters.

Figure 6 are plots when $n = 50$, Figure 7 are plots when $n = 500$, and Figure 8 is a partial plots of Fig.2 showing more details of low degrees. For both $n = 50$ and $n = 500$, the tails of distributions are similar, partly due to the small amount of data in these tails. When removing more and more vertices, the expected degree distribution becomes more away from the original degree distribution. However we cannot say that it is more away from power-law distribution, since it might be fitted by other power-law distribution with new parameters.

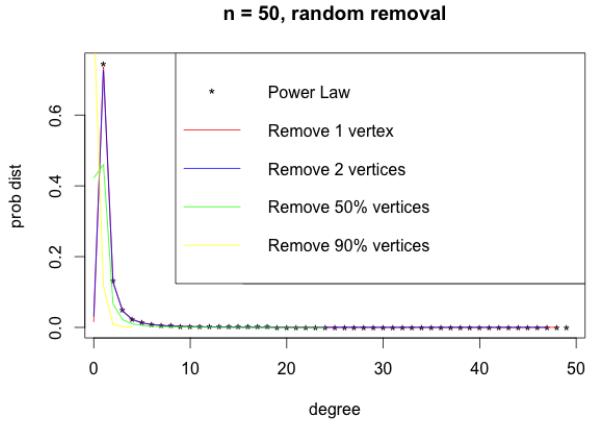

Figure 6: Plots of the original power-law degree distribution and expected degree distribution after different proportions of removals: remove one vertex, remove two vertices, remove 50% of the vertices, and remove 90% of the vertices; n=50.

4.3.2.3 Normal curve removal: One vertex removed. For random removal, as shown above, the expected degree distribution does not use the information in$W$. However, following the same

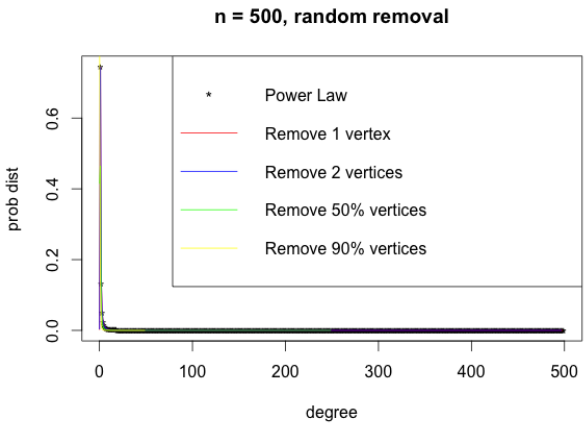

Figure 7: Plots of the original power-law degree distribution and expected degree distribution after different proportions of removals; n=500.

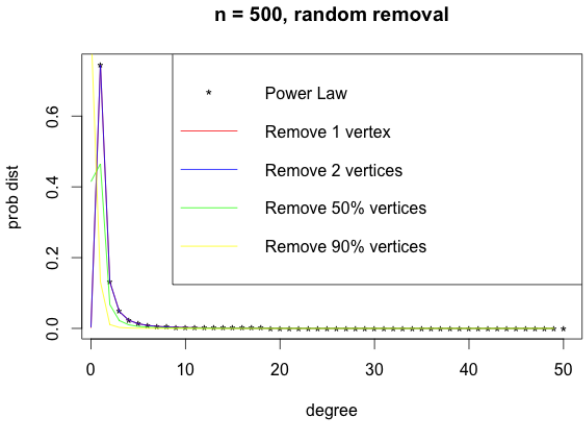

Figure 8: This figure shows a part of Figure 2 for better viewing of details for degrees at most 50.

approach to derive from (4.1), the interchange of summation in (4.3), which is available for random removal because of uniform removal probabilities, is not available for normal curve removal. Therefore, we tried another perspective to calculate q_k; recall that

$$q_k = \mathrm{E}(q_k^{T_M,M,W}) = \mathrm{E}(\mathrm{E}(q_k^{T_M,M,W})|M) = \sum_{m=1}^{n-1} P(M=m)E(q_k^{T_m,m,W}|M=m), \tag{4.8}$$

where $E(q_k^{T_m,m,W}|M=m)$ is an expected proportion. We wrote this proportion by its definition as the ratio of number of vertices with degree k over $n-1$. Just as for random removal, there are two types of vertices with degree k after removal. Some are vertices with degree k before removal and without a connection with the removed vertex; the others are vertices with initial degree $k+1$, and are connected with the removed vertex, so they lose one degree upon removal. We denote number of vertices with degree k which are removed together with the removed vertex $v^{T_m,m,W}$ as $a_k^{T_m,m,W}$ (note that here $a_k^{T_m,m,W}$ only includes number of vertices connected with the removed vertices, not of the removed vertices with degree k) to get

$$E(q_k^{T_m,m,W}|M=m) = E\left(\frac{np_k - a_k^{T_m,m,W} - \mathbb{1}_{m=k} + a_{k+1}^{T_m,m,W}}{n-1}|M=m\right) \tag{4.9}$$

We propose Assumption 1 to facilitate the calculation of (4.9). When a vertex v with degree m is removed, knowing $\mathrm{E}(a_k^{T_m,m,W})$, or on average how many vertices with degree k lose one edge together with v, is equivalent to knowing what proportion of the m removed edges of v are connected with the vertices with degree k. Assumption 1 provides this proportion. Multiplying that by m yields the average number of vertices with degree k losing one edge due to removal.

Assumption 1 For any vertex in a network, the expected proportion of the edges connected to vertices with degree k in all edges of this vertex, is the proportion of the degrees of vertices with degree k in all degrees of all vertices. Also, assume that for any vertex $v_0 \in V$,

$$\frac{d(v_0)}{\sum_v d(v)} < \frac{1}{\max\limits_{v\in V}(d(v))},$$

that the network is a simple network, which is undirected, without multiple edges or loops.

We use a social network as an example to illustrate Assumption 1. Suppose Mike has 500 friends in a social network (consider Mike a vertex in this network). How many of his friends have 100 friends ($k=100$)? Suppose that in this network, there are in total 200 people whose number

of friends is 100, then total friends number of these people is $200 * 100 = 20,000$ (for simplicity we do not consider overlap of friends among these 200 people). Suppose the total number of friends of all people in this network is $400,000$. Then Assumption 1 states that the proportion of 100-friend people among Mike's friends is the proportion of those people's friends in all people's friends, which is $200 * 100/400,000 = 5\%$. Therefore, Mike should have $500 * 5\% = 25$ friends whose number of friends is 100. $\frac{d(v_0)}{\sum_v d(v)} < \frac{1}{\max\limits_{v \in V}(d(v))}$ indicates that even the most popular person (who has most friends) will not be any person's friend twice, which guarantees that there are no multiple edges between any two persons.

For many network generation schemes, the generation procedure depends only on degrees of vertices, treating each single link with equal weight. Such generation procedures include configuration modeling, B-A model, Chung model, and others (Newman, 2005). Therefore, it is reasonable to propose Assumption 1, which assumes properties of connections of a vertex depends only on degrees of vertices connected to it, with each link equally weighted.

With Assumption 1, we have

$$E(a_k^{T_m,m,W}|M=m) = m\frac{np_k k}{\sum_{l=1}^{n-1} np_l l} \quad \text{and} \quad E(a_{k+1}^{T_m,m,W}|M=m) = m\frac{np_{k+1}(k+1)}{\sum_{l=1}^{n-1} np_l l};$$

therefore,

$$(4.9) = \frac{m\frac{np_{k+1}(k+1)}{\sum_{l=1}^{n-1} np_l l} + np_k - m\frac{np_k k}{\sum_{l=1}^{n-1} np_l l} - \mathbb{1}_{m=k}}{n-1} \tag{4.10}$$

Now plug in (4.10) into (4.8) to get the expected degree distribution:

$$\begin{aligned} q_k &= \sum_{m=1}^{n-1} np_m \pi(v^m) \frac{m\frac{np_{k+1}(k+1)}{\sum_{l=1}^{n-1} np_l l} + np_k - m\frac{np_k k}{\sum_{l=1}^{n-1} np_l l} - \mathbb{1}_{m=k}}{n-1} \\ &= \sum_{m=1}^{n-1} np_m \pi(v^m) \frac{m\frac{nC(k+1)^{-\alpha}(k+1)}{\sum_{l=1}^{n-1} nCl^{-\alpha} l} + nCk^{-\alpha} - m\frac{Ck^{-\alpha} k}{\sum_{l=1}^{n-1} nCl^{-\alpha} l} - \mathbb{1}_{m=k}}{n-1} \end{aligned} \tag{4.11}$$

Here v^m is any vertex with degree m, and $\pi(v^m)$ depends only on m. (4.9) is also a linear combination of $k^{-\alpha}$ and $(k+1)^{-\alpha}$, which is not a power-law distribution. Therefore, for normal curve removal, power-law distribution is also destroyed when just one vertex is removed.

In fact, we notice that during the derivation of the normal curve removal, we used a general notation $\pi(v)$ to denote the removal probability rather than a specific normal form. Therefore, the

derivation is a general one which can be used for all removal strategies. We check its correctness by applying it to the random removal. Using $\pi(v^m) = 1/n$ in line 1 of (4.11), we see that (4.12) is same as (4.3), as expected:

$$\begin{aligned}
(4.9) &= \frac{1}{n(n-1)}\left(\frac{p_{k+1}(k+1)-p_k k}{\sum_{l=1}^{n-1} l p_l l}\sum_{m=1}^{n-1} np(m)m + n^2 p_k - np_k\right) \\
&= \frac{1}{n(n-1)}\left(\frac{p_{k+1}(k+1)-p_k k}{\sum_{l=1}^{n-1} p_l l}\sum_{l=1}^{n-1} np_m m + n^2 p_k - np_k\right) \\
&= \frac{1}{n-1}(p_{k+1}(k+1) + (n-(k+1))p_k)
\end{aligned} \tag{4.12}$$

Two and more vertices removed

Assumption 1 provides a way to calculate the expected degree distribution for the normal curve removal when one vertex is removed. However, for the normal curve the removal of more than one vertex is more complicated, for it requires more information or assumptions. The formula of the expected degree distribution when two vertices are removed is given in (4.13). Abbreviating the removed two vertices $v_1^{T_{m_1},m_1,W}$ and $v_1^{T_{m_2},m_2,W}$ as v_1 and v_2, we have

$$\begin{aligned}
q_k &= \mathrm{E}(\mathrm{E}(q_k^{T_{m_1},T_{m_1},m_1,m_2,W}|M_1,M_2)) \\
&= \sum_{m_1=1}^{n-1}\sum_{m_2=1}^{n-1} P(M_1=m_1,M_2=m_2)\mathrm{E}\Big(\frac{\sum_{v\in V}(\mathbb{1}_{d(v)=k,l(v,v_1)=0,l(v,v_2)=0} + \mathbb{1}_{d(v)=k+1,l(v,v_1)=1 \mathrm{or} l(v,v_2)=1}}{n-2} \\
&\quad + \frac{\mathbb{1}_{d(v)=k+2,l(v,v_1)=1,l(v,v_2)=1}\big)}{n-2}\Big)
\end{aligned} \tag{4.13}$$

Actually, the expected value of the summation of each indicator function in (4.13) is not easy to calculate. Consider $\mathrm{E}(\sum_{v\in V}\mathbb{1}_{d(v)=k+2,l(v,v_1)=1 \mathrm{and} l(v,v_2)=1})$ for an example, which is to calculate averagely how many vertices with degree $k+2$ will lose 2 edges during the removal of v_1 and v_2, one edge from each. Note that there is no information about how many links between v_1 and v_2, such that it is uncertain that how many edges are removed totally by v_1 and v_2. What is more, even the number of links between v_1 and v_2 is known, then suppose there are totally m edges removed from the unit of v_1 and v_2, we still cannot use Assumption 1. Because each of the connected vertices with degree $k+2$ is supposed to connect with the unit twice, while Assumption 1 is for simple network where double linkages are not allowed. Therefore, there is no further derivation for Eq 4.13.

4.3.2.4 High degree removal. Since the derivation for the normal curve removal is actually general for all removal strategies, we can directly apply it to the high degree removal. Above, we considered removing only one vertex. The removal probability $\pi(v) = 1$ if v is with highest degree (or say, largest degree), where h is number of highest degree, and $\pi(v) = 0$ if v is not with highest degree. Plugging in $\pi(v)$ into (4.12) and (4.13), we can see that the expected degree distribution is not a power-law distribution. Also, we are not able to get the formula for the cases when more than two vertices are removed because the expressions are intractable.

4.3.3 Experiment 2

Although we showed theoretically that the power-law distribution will not hold after just one vertex is removed for all attack strategies, Experiment 1 shows that empirically, for the random removal and normal curve removal, the degree distribution is very resilient, that is, close to a power-law distribution. Such a discrepancy is often tolerable in applications. The degree distribution may well retain power-law properties if it is perturbed a bit from a power-law distribution, because those properties are very useful. Under such cases, if the power-law distribution is still considered valid, parameters might change. In Experiment 2, we use a simulation study to explore whether and how the parameters of the degree distribution will change when power-law distribution is considered valid after a goodness of fit test.

We still use the same three networks as those in the Experiment 1, and for each network, we plot the mean and standard deviation of estimates of γ and α. We used both random attack and normal curve attack strategies, which show power-law resilience in the experiment 1. Results are shown in Figure 9 and Figure 10 respectively.

From Figure 9 and Figure 10, we can conclude the following.

1. The trends in changes of the parameters of all three networks look similar for each strategy.
2. Though the change trends are similar, the B-A network appears to have smaller standard deviations than the other two networks for both α and γ. Protein network shows a change of parameters earlier than the other two networks for both α and γ.
3. For both the random and normal curve removals, values of α are quite stable, staying around the original value with small standard deviations. For random removal, α retains its value until

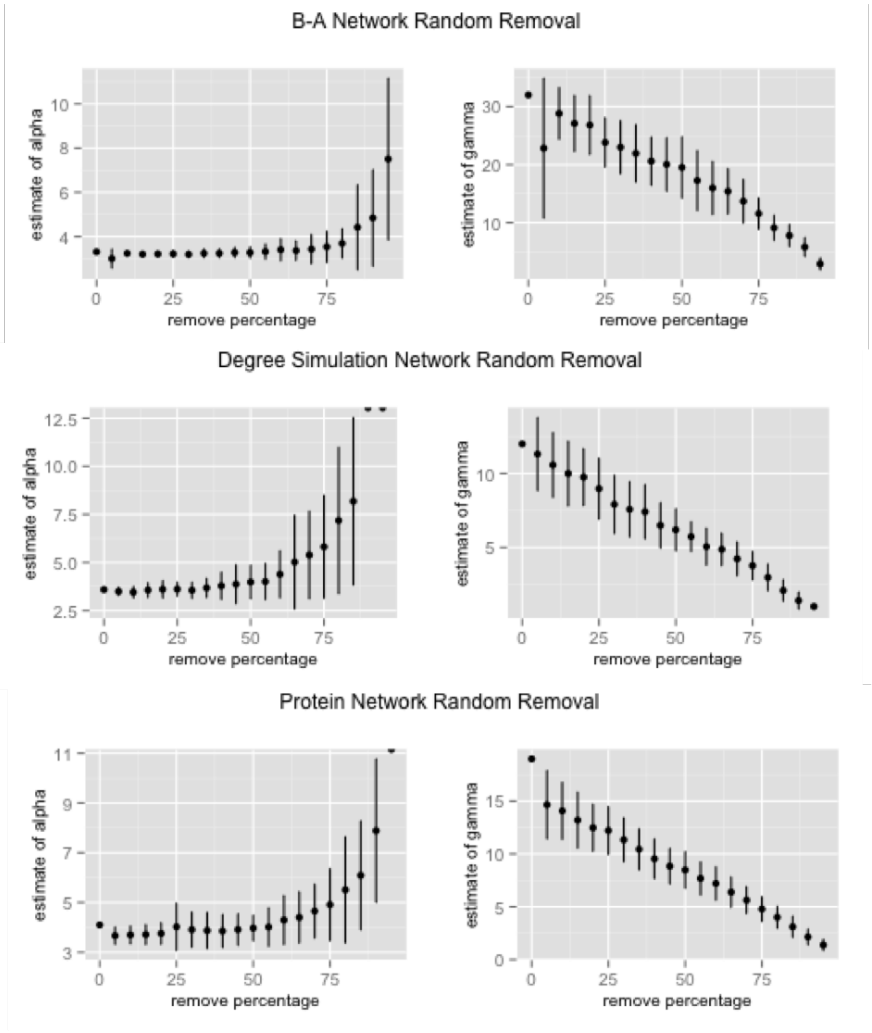

Figure 9: Estimation of parameters of power-law part of the degree distribution after random removal of vertices. We remove vertices from 5% to 95%, in steps of 5% each time. For each network, we repeat the removing procedure 50 times to calculate the mean and standard deviation of parameter estimates.

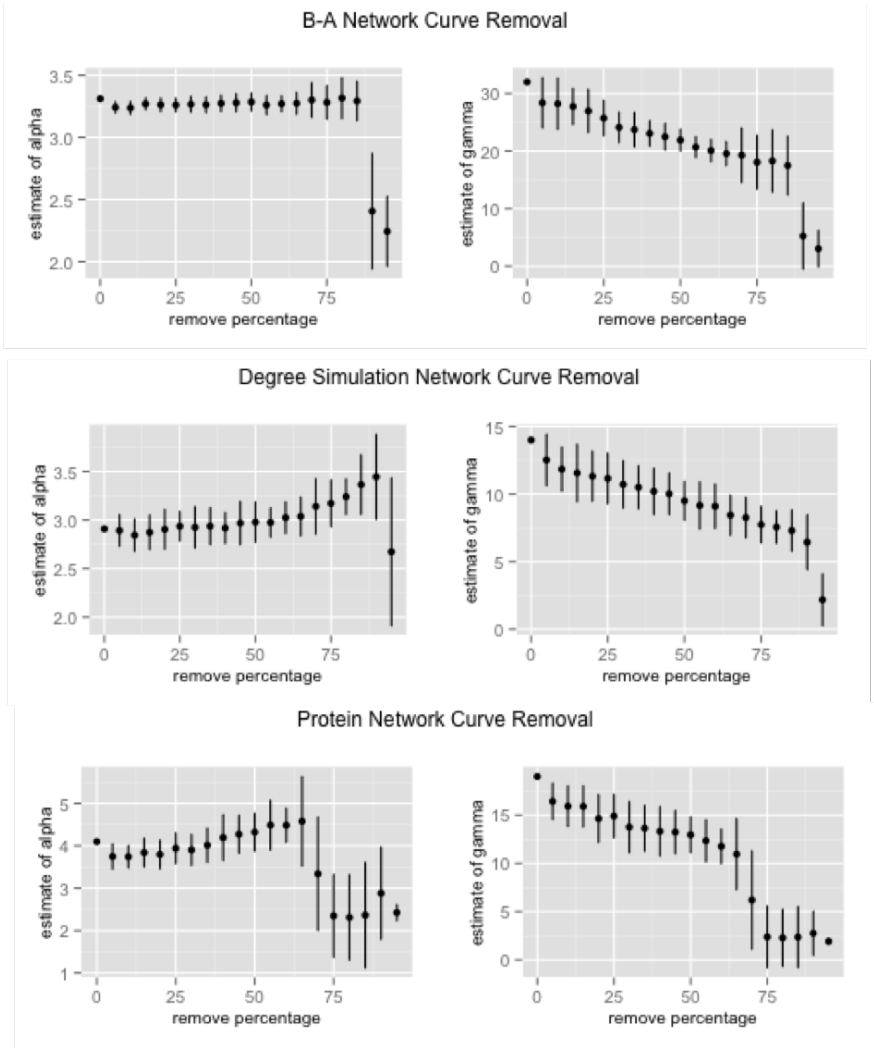

Figure 10: Estimation of parameters of power-law part of the degree distribution after normal curve removal of vertices. The only difference between Fig 9 is that the removal strategy here is normal curve removal.

more than 75% vertices removed. After that, values increase and standard deviations increase. For normal curve removal, α stays stable until more than 75% vertices removed too; after that, values decrease and standard deviations also increase.

4. For both random removal and normal curve removal, values of γ decrease. For random removal, values of γ decrease continuously, with standard deviation starting to decrease when around 60% vertices are removed. For normal curve removal, values of γ decrease continuously first, while it appears to have a sudden decrease when around 75% of the vertices are removed. Standard deviations of γ in normal curve removal appears to be unchanged (though B-A network appears smaller standard deviations in the middle part).

4.4 SUMMARY

- We conducted a simulation study to see whether the degree distribution is still a power-law distribution after attacks by three removal strategies.
- We derived the mathematical form of the expected degree distribution for three removal strategies. For random removal, we derived the form when any number of vertices are removed. For normal curve and high degree removals, we derived it when only one vertex is removed, under certain assumptions. We also pointed out that the derivation for normal curve removal and high degree removal is not tractable when more than two vertices are removed.
- We conducted a simulation study to see how parameters change if the power-law degree distribution is considered valid after a goodness of fit test.

BIBLIOGRAPHY

Aban, I. B., Meerschaert, M. M., and Panorska, A. K. (2006). Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473):270–277.

Apellániz, J. M. and Úbeda, L. (2005). Numerical biases on initial mass function determinations created by binning. *The Astrophysical Journal*, 629(2):873.

Barabási, A.-L. et al. (2009). Scale-free networks: a decade and beyond. *science*, 325(5939):412.

Bauke, H. (2007). Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B*, 58(2):167–173.

Beg, M. (1983). Unbiased estimators and tests for truncation and scale parameters. *American Journal of Mathematical and Management Sciences*, 3(4):251–274.

Beirlant, J., Alves, I. F., and Gomes, I. (2015). Tail fitting for truncated and non-truncated pareto-type distributions. *arXiv preprint arXiv:1505.05189*.

Brazauskas, V. and Serfling, R. (2003). Favorable estimators for fitting pareto models: A study using goodness-of-fit measures with actual data. *Astin Bulletin*, 33(02):365–381.

Burroughs, S. M. and Tebbens, S. F. (2001). Upper-truncated power law distributions. *Fractals*, 9(02):209–222.

Callaway, D. S., Newman, M. E., Strogatz, S. H., and Watts, D. J. (2000). Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468.

Clark, D. R. (2013). A note on the upper-truncated pareto distribution. In *Proc. of the Enterprise Risk Management Symposium, April*, pages 22–24.

Clarke, B., Fokoue, E., and Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer Science & Business Media.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Cohen, R., Erez, K., Ben-Avraham, D., and Havlin, S. (2000). Resilience of the internet to random breakdowns. *Physical review letters*, 85(21):4626.

Cohen, R., Erez, K., Ben-Avraham, D., and Havlin, S. (2001). Breakdown of the internet under intentional attack. *Physical review letters*, 86(16):3682.

Coile, R. C. (1977). Lotka's frequency distribution of scientific productivity. *Journal of the american society for information science*, 28(6):366–370.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis*, 76(2):226–248.

Deluca, A. and Corral, Á. (2013). Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophysica*, 61(6):1351–1394.

Goldstein, M. L., Morris, S. A., and Yen, G. G. (2004). Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41(2):255–258.

Gutenberg, B. and Richter, C. F. (1944). Frequency of earthquakes in california. *Bulletin of the Seismological Society of America*, 34(4):185–188.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of multivariate analysis*, 32(2):177–203.

Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nature biotechnology*, 23(7):839–844.

Hill, B. M. et al. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 3(5):1163–1174.

Holme, P., Kim, B. J., Yoon, C. N., and Han, S. K. (2002). Attack vulnerability of complex networks. *Physical Review E*, 65(5):056109.

Koen, C. (2006). On the upper limit on stellar masses in the large magellanic cloud cluster r136. *Monthly Notices of the Royal Astronomical Society*, 365(2):590–594.

Kolaczyk, E. D. (2009). Sampling and estimation in network graphs. In *Statistical Analysis of Network Data*, pages 1–30. Springer.

Maschberger, T. and Kroupa, P. (2009). Estimators for the exponent and upper limit, and goodness-of-fit tests for (truncated) power-law distributions. *Monthly Notices of the Royal Astronomical Society*, 395(2):931–942.

Neukum, G. and Ivanov, B. (1994). Crater size distributions and impact probabilities on earth from lunar, terrestrial-planet, and asteroid cratering data. *Hazards due to Comets and Asteroids*, 359.

Newman, M. E. (2002). Spread of epidemic disease on networks. *Physical review E*, 66(1):016128.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.

Nuyts, J. (2010). Inference about the tail of a distribution: Improvementon the hill estimator. *International Journal of Mathematics and Mathematical Sciences*, 2010.

Pareto, V. (1964). *Cours d'économie politique*. Librairie Droz.

Peters, O., Deluca, A., Corral, Á., Neelin, J. D., and Holloway, C. E. (2010). Universality of rain event size distributions. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(11):P11030.

Quandt, R. E. (1966). Old and new methods of estimation and the pareto distribution. *Metrika*, 10(1):55–82.

Singh, D., Prakash, G., and Singh, P. (2007). Shrinkage testimators for the shape parameter of pareto distribution using linex loss function. *Communications in Statistics—Theory and Methods*, 36(4):741–753.

Stumpf, M. P., Wiuf, C., and May, R. M. (2005). Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4221–4224.

White, E. P., Enquist, B. J., and Green, J. L. (2008). On estimating the exponent of power-law frequency distributions. *Ecology*, 89(4):905–912.